MODERN INFORMATION RETRIEVAL SYSTEMS

BS(LIS)

Code No. 9214 **Units: 1-9**

Department of Library and Information Sciences
Faculty of Social Sciences and Humanities
ALLAMA IQBAL OPEN UNIVERSITY
ISLAMABAD

MODERN INFORMATION RETRIEVAL SYSTEMS

BS (LIBRARY AND INFORMATION SCIENCES)

Course Code: 9214 **Units: 1–9**

Department of Library and Information Sciences
Faculty of Social Sciences and Humanities
Allama Iqbal Open University
Islamabad

First Printing 2022

Quantity..... 1000

Layout Setting..... Ikram Yousaf

Incharge Printing..... Dr. Sarmad Iqbal

Printer..... AIOU-Printing Press, H-8, Islamabad

Publisher ..... Allama Iqbal Open University, Islamabad

COURSE TEAM

Chairman Course team:

Prof. Dr. Pervaiz Ahmad

Chairman

Course Development Coordinator:

Dr. Munazza Jabeen

Assistant Professor

Compiled by:

Dr. Munazza Jabeen

Reviewed by:

Dr.Muhammad Arif

Members

Dr. Pervaiz Ahmad

Dr. Muhammad Arif

Dr. Munazza Jabeen

Dr. Amjad Khan

Muhammad Jawwad

FOREWORD

Department of Library and Information Sciences was established in 1985 under the flagship of the Faculty of Social Sciences and Humanities intending to produce trained professional manpower. The department is currently offering seven programs from certificate course to PhD level for fresh and/or continuing students. The department is supporting the mission of AIOU keeping in view the philosophies of distance and online education. The primary focus of its programs is to provide a quality education through targeting the educational needs of the masses at their doorstep across the country.

BS 4-year in Library and Information Sciences (LIS) is a competency-based learning program. The primary aim of this program is to produce knowledgeable and ICT-based skilled professionals. The scheme of study for this program is specially designed on the foundational and advanced courses to provide in-depth knowledge and understanding of the areas of specialization in librarianship. It also focuses on general subjects and theories, principles, and methodologies of related LIS and relevant domains.

This new program has a well-defined level of LIS knowledge and includes courses of general education. The students are expected to advance beyond their higher secondary level and mature and deepen their competencies in communication, mathematics, languages, ICT, general science, and array of topics social science through analytical and intellectual scholarship. Moreover, the salient features of this program include practice-based learning to provide students with a platform of practical knowledge of the environment and context, they will face in their professional life.

This program intends to enhance students' abilities in planning and controlling library functions. The program will also produce highly skilled professional human resources to serve libraries, resource access centers, documentation centers, archives, museums, information centers, and LIS schools. Further, it will also help students to improve their knowledge and skills of management, research, technology, advocacy, problem-solving, and decision-making relevant to information work in a rapidly changing environment along with integrity and social responsibility. I welcome you all and wish you good luck with your academic exploration at AIOU!

Prof. Dr. Zia Ul-Qayyum
Vice-Chancellor

PREFACE

Traditionally, information retrieval has concerned only a selected few, specifically people engaged in knowledge-intensive activities such as education and research. Therefore, it required access to data in various forms (primarily digital). However, in today's world, we use information retrieval systems in almost every aspect of our daily lives: retrieving an e-mail message received or sent on a specific data to a specific person; finding something or someone on the web; searching for a book in an online library catalogue or a digital library; searching for a song or finding a video on YouTube; and so on. This expansion of information retrieval in specific knowledge-intensive and daily activities has brought new challenges and opportunities. That has resulted in a tremendous amount of research and development activities in information retrieval. While the original aim of this study guide is to provide a blend of traditional and new approaches to information retrieval. The study guide endures covering the whole spectrum of information storage and retrieval on away relevant to a specific librarianship perspective.

The primary audience comprises Library and Information Science Programs students, both at undergraduate and postgraduate levels. This study guide has been organized into nine units and aims to help students complete the required coursework. Each unit starts with an introduction that provides an overview, followed by objectives that explain how a student, after reading the unit, students should be able to describe, compare, and analyze the concepts studied in that unit. Hence, this study guide intends to be a concise document and learning tool in which the course contents are briefly introduced. It also strives to make the students knowledgeable and successful in /her endeavours. The study guide provides a list of books & suggested reading based on the course requirement. This study guide is expected to meet the requirements of students undertaking course information retrieval systems.

However, it will help practising library and information professionals to brush up on their knowledge in different areas of information retrieval. I congratulate the course development team for their hard work and commitment to completing this study guide.

Prof. Dr. Syed Hassan Raza
Dean, Faculty of Social Sciences and Humanities

ACKNOWLEDGMENTS

We are extremely grateful to the worthy Vice-Chancellor and Dean, Faculty of Social Sciences and Humanities for us the opportunity to prepare this study guide for the BS (library & information science) program. With their kind support, this task has been made possible.

Special thanks to the Academic Planning and Course Production and Editing Cell of AIOU for their valued input to improve the quality of this study guide. We also thank the Print Production Unit of the University for the formatting of the manuscript and final production. We also appreciate the efforts of ICT officials, the staff of the central library, and the LIS department to accomplish this academic task. In the end, we also appreciate the extended cooperation of the course team in this academic task.

Dr. Munazza Jabeen

TABLE OF CONTENTS

Page #

INTRODUCTION

An information retrieval (IR) system is a set of algorithms that facilitate the relevance of displayed documents to searched queries. In simple words, it works to sort and rank documents based on the queries of a user. There is uniformity with respect to the query and text in the document to enable document accessibility. The study guide aims to cover the whole spectrum of information storage and retrieval in a way that is relevant to students.

Unit 1 has expressed not only the conceptual framework but also the perspectives of information retrieval. It has provided insight on from print to web information resources. Unit 2 has provided discussions on some bibliographic formats in favor of new sections on the MARC machine-readable cataloguing) format. Unit 3 significantly discussed the recent developments in online public access catalogues, and new topics, such as cataloguing of internet resources and functional requirements for bibliographic records (FRBR). Moreover, it discusses the issues of vocabulary control in information retrieval. Unit 4 deals with basic concepts of the information search process along with various information retrieval models. It identified a wide range of search strategies as well. It has included some discussions of extensible markup language (XML) retrieval. Unit 5 covers the issues related to information users and the various approaches to user studies like user-centered information retrieval models. Unit 6 discusses the possibilities of online and CD-ROM information retrieval. Furthermore, it has elaborated the new features of online database search services. Unit 7 discussed markup languages in the context of information retrieval. Unit 8 concentrates on natural language processing in an information retrieval system. Unit 9 discusses various aspects of information retrieval in digital libraries; it has provided extensive updates on recent developments and examples such as sophisticated information retrieval applications in databases and search engines, and social information retrieval.

OBJECTIVES

After studying this course, modern information retrieval systems, you should be able to comprehend the following concepts:

- Information retrieval systems
- Concepts about online cataloging and metadata
- Indexing and abstracting
- User interfaces
- Evaluation of information retrieval systems
- Hypertext and markup language and web information retrieval
- Information retrieval in digital libraries
- Trends in information retrieval

Unit–1

BASIC CONCEPTS OF INFORMATION RETRIEVAL SYSTEMS, DATABASE TECHNOLOGY, AND BIBLIOGRAPHIC FORMATS

Complied by: Dr. Munazza Jabeen

Reviewed by: 1. Dr. Pervaiz Ahmad
2. Dr. Muhammad Arif
3. Dr. Amjid Khan

CONTENTS

INTRODUCTION

Database technology emerged in the late sixties as a result of a combination of various circumstances. There was a growing demand among users for more information to be provided by the computer relating to the day-to-day running of the organization as well as to planning and control purposes. This demand coincided with advances in computer technology and in expertise in the computer data processing. The technology that emerged to process and manipulate data of various kinds is broadly termed 'database management technology', and the resulting software packages are known as database management systems (DBMSs). DBMSs do just what the name suggests: they manage a computer-stored database or collection of data.

Basic concepts of database systems, their growth, and recent trends in database technology are discussed in this unit. Our primary concern is with bibliographic or text databases which form the basis of information retrieval systems. Different kinds of bibliographic/text databases are mentioned by way of examples. Finally, measures to be taken to develop databases in an information retrieval environment are briefly discussed.

An information retrieval system should create and maintain one or more databases containing records pertaining to the requirements of the user community. In any organization, different kinds of information may be required. A large proportion of information required is factual: the contents of the database — the records, contain various facts such as the features of a particular chemical element or compound, a metal, a tool, a piece of equipment, an automobile, a spare part, a drug, a patient, a plant, a forest, an agrochemical, a national park, and so on. The creation and maintenance of such a factual information retrieval system require background knowledge of (i) the subject field and (ii) the actual and potential users and their activities vis-a-vis their information requirements and interests. Such a database system is usually developed for use by people within an organization/institution or in a group of organizations/institutions, but the data are not expected to be accessible to everyone as happens in the case of library databases. Decisions relating to the database structure, format, and data exchange mechanism are governed in such cases by several factors, such as the chosen database management software, the database design principle, and moreover the needs and access rights of the user community.

OBJECTIVES

After reading this unit, you would be able to:

1. Information about re-searching information retrieval systems
2. To know about types of bibliographic databases
3. Knowledge about in-house and online IRS software
4. Generating indexes for efficient information retrieval
5. Database design

1.1 FEATURES OF AN INFORMATION RETRIEVAL SYSTEM

Figure 1.1 presents the conceptual view of an information retrieval system. An information retrieval system is designed to enable users to find relevant information from a stored and organized collection of documents. Thus, the concept of information retrieval presupposes that there are some documents or records containing information that has been organized in an order suitable for easy retrieval.

1.2 ELEMENTS OF N INFORMATION RETRIEVAL SYSTEM

Figure 1.1 shows that an information retrieval system for social sciences has one or more different types of documents and can contain text as well as multimedia information. All the documents are processed to create an index, which is searched for retrieval of information.

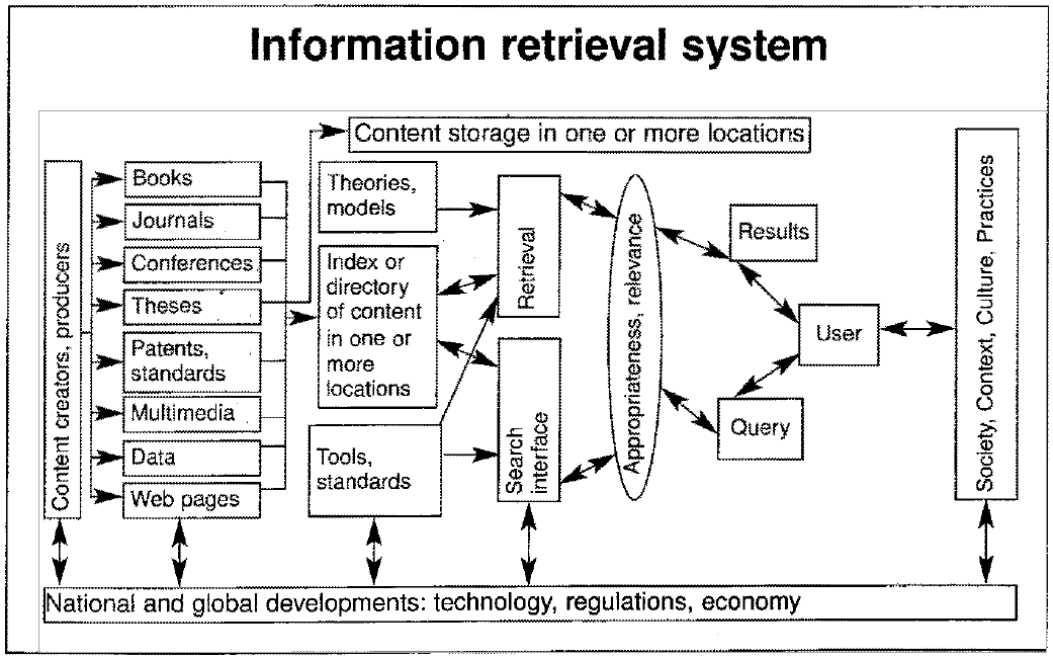

Figure 1.1 Broad outline of an IRS

1.3 PURPOSE

An information retrieval system is designed to retrieve the documents or information required by the user community. It should make the right information available to the right user. Thus, an information retrieval system aims to collect and

organize information in one or more subject areas to provide it to users as soon as they ask for it. Belkin4 describes how information retrieval systems are used.

1.4 FUNCTIONS

An information retrieval system deals with various sources of information on the one hand and users' requirements on the other. It must:

- Analyses the contents of the sources of information as well as the users' queries, and then match these to retrieve those items that are relevant.

1.5 COMPONENTS

It is evident from the above discussion that on the one side of an information retrieval system there are the documents or sources of information and on the other there are the users' queries. These two sides are linked through a series of tasks. Lancaster mentions that an information retrieval system comprises six major subsystems:

- the document subsystem
- the indexing subsystem R the vocabulary subsystem R the searching subsystem
- the user-system interface
- the matching subsystem.

1.6 KINDS OF INFORMATION RETRIEVAL SYSTEMS

Information retrieval systems can be categorized in several ways. For example, one can group them into two categories: in-house and online. In-house information retrieval systems are set up by a particular library or information center to serve mainly the users within the organization. One in-house database is the library catalogue. OPACs provide facilities for library users to carry out online catalog searches and to then check the availability of the item required.

1.7 DESIGN ISSUES

A system can be defined as a set of interacting components, under human control, operating together to achieve an intended purpose. Thus, a system carries out processing on inputs to produce required outputs; the agents of this processing are people and machines.

1.8 DISCUSSION

Developments in information retrieval can be viewed from two different perspectives:

- the computer-centered view, which deals with building efficient computer systems for storage, organization, and access to information, and focuses on areas such as building up efficient access mechanisms, query processing, ranking algorithms and display and delivery of search results
- The user-centered view, focuses on the study of human information behavior, understanding of human needs, information context and use, and so on.

1.9 DATA

The word 'data' refers to a set of given facts. Information in a form that can be processed by a computer is called data. Data has for a long time been used to refer to scientific measurements, but words constitute data just as numbers do. A list of names is data, a set of keywords is data, a doctor's record of their patients is data, and figures relating to temperature, humidity, and so forth, or sales of a company, are data.

The Database

A database can be conceived as a system whose base, whose key concept, is simply a particular way of handling data. In other words, a database is nothing more than a computer-based record-keeping system. The overall objective of a database is to record and maintain information. The Macmillan Dictionary of Information Technology defines a database as a collection of interrelated data stored so that it may be accessed by users with simple user-friendly dialogues'. The Chambers Science and Technology Dictionary provides a simpler definition of a database: 'a collection of structured data independent of any particular application'.

Records and Fields

A *record* is a collection of related information. A database is an organized collection of units of information, and each unit of information in a database is called a record. A record is generally what a user wants to find while searching a database. An example of a record is a book card in a library's catalog, which describes the book's title, author, subject, and so on. A collection of database records constitutes a database file. Identifying what the record is to be is one of the early tasks in designing a database. If the database is a bibliographic one, the bibliographic information about each document is the unit of information or record.

Properties of Database

A database is designed to avoid duplication of data as well as to permit retrieval of information to satisfy a wide variety of user information needs. Major properties of a database can be summarized as follows:

- it is integrated with provisions for different applications
- it eliminates or reduces data duplication
- it enhances data independence by permitting application programs to be insensitive to changes in the database

Kinds of Databases

In discussing databases, it is sometimes useful to classify them by the type of data record contained and sometimes by subject coverage. The two major divisions are reference databases and source databases. Reference databases lead the users to the source of the information: a document, person, or organization. They can be divided into three categories:

- The bibliographic databases, which include citations or bibliographic references, and sometimes abstracts of literature
- The catalog databases, which show the catalogue of a given library or a group of libraries in a network, and
- The referral databases, which offer references to information such as the name, address and specialization of persons, institutions, information systems, and so on.

Database Technology

The historical development of database technology has been closely related to the development of computer hardware and software. With respect to hardware development, it is now common to talk about 'computer generations', and in a similar way several 'database system generations' can be distinguished.

The Development of Databases in An Information Retrieval Environment

An information retrieval system may contain various kinds of databases. The data may be factual, containing information required for research, planning, management, and for all kinds of day-to-day activities. Such databases may include, for instance: in a healthcare information system, information related to drugs, patents and so forth; in a pollution-control environment, information related to various chemicals, pollutants, plants, parks, and so forth; in a forestry-management environment, forests, plants and so on; in an automobile information system, vehicles, spare parts, and so on.

Basic Considerations

In most general terms, to run an information retrieval system we need the following:

- a software (text retrieval) package
- a processor to execute the programs
- memory to hold intermediate working
- disk storage to hold the data files
- devices for archiving data files to recover from accidental damage or loss of data
- printer(s) to produce hard copy for different purposes, and
- terminals for data input and for controlling the whole process.

Data Base Design

Designing the database constitutes the first step of developing a text retrieval system. This step involves a number of important decisions and the performance of the resulting system will depend largely on these. A text retrieval system may be used for a variety of operations — preparation of library catalogues, bibliographies, current awareness lists, biographical lists, and so on. In each of these cases, the nature of the database and consequently the nature, content and number of fields will differ.

Database Indexing

This is an important step in any text retrieval system because it will generate the index file on which searches can be performed. Most text retrieval systems create an inverted index file. Software packages have different mechanisms to indicate which of the fields should be indexed and how this should be done, and a database designer must follow those steps. In some software packages, the index file is generated and updated as soon as new records are added or existing records are deleted, while in others, the creation and update dating index file have to be done by a batch modes again this process is software dependent.

Data Entry Form/Worksheet

Another important task involved in the database design stage relates to the creation of the data entry form or worksheet, which is like a blank form used for entering data in the database. In some text retrieval packages, the designer has to create a data entry worksheet which is used by the data entry operators to enter data into the database. However, some text retrieval software allows data to be entered directly without the need to create any form or worksheet for the purpose.

Output Format

The database designer has to consider how the user will expect the records to be displayed when the database is browsed or when records are retrieved by a certain search. Some software allows the designer to produce one or more output formats for displaying the retrieved records. However, the facility of displaying records in a chosen format is not available in all text retrieval systems.

Data Entry, Searching, and Printing

The job of database design ends with the tasks mentioned above. The next job is the creation of records. This involves entering data elements in the appropriate columns in the worksheet or form for data entry. This can be done in one of two ways. Records can be created by keying in the data elements in each field and subfield in the data entry form/worksheet or a number of records can be downloaded from other, already existing, databases.

1.10 DISCUSSION

In the previous section, the various steps that one has to follow to develop a database using text retrieval software were described. The specific steps and measures prescribed for each operation differ from program to program, but some points may be generalized. With these basic considerations in mind, one must follow the specific prescriptions of the chosen software. The major issues involved here are:

- design of the database structure
- decisions regarding the generation of the index file
- decisions regarding the format of data display
- design of the worksheet or form for data entry
- creation of records
- generation of the index file
- searching the database, and
- displaying, sorting, and printing records.

1.11 BIBLIOGRAPHIC RECORDS

The term ‘bibliographic record’ is relatively new, having entered the information vocabulary mainly as a result of automation. It has been defined as ‘the sum of all the area and elements which may be used to describe, identify or retrieve any physical item (publication, document) of information content.

1.12 BIBLIOGRAPHIC FORMATS

Different data elements are very closely tied up with content designators. The data elements separately identified by the codes in the exchange formats have to be defined, not only in terms of content but also in form, if the records are to be suitable for use by another agency. Effective exchange of bibliographic data between agencies can be accomplished only if the records of agencies exchanging data

conform in respect of all the three components: the structure, the content designators, and the data element definitions.

1.13 ACTIVITIES

1. Investigate the conceptual view of an information retrieval system to retrieve documents you want to implement for your IRS assignment?
2. Search the availability of information retrieval systems in your area? How will you generate an index for minimizing response time in searching the documents for your IRS project?
3. Evaluate various database categories to classify the reference or source of the information. Which solution is more suitable for your IRS project? Hint: compare the objectives of your IRS with the pros and cons of bibliographic, catalog, and referral databases.
4. Design your information retrieval system. Which options will work best for your system in terms of displaying, sorting, and printing records?

1.14 SELF-ASSESSMENT QUESTIONS

1. How searching for bibliographic records can benefit from an information retrieval system?

2. Explain the difference between in-house and online IR systems.

3. How subscription of software (text retrieval) package is different from licensing and hosted systems?

4. How can archiving data files help to recover from accidental damage or loss of data?

1.15 REFERENCES

Belkin, N. J. (1980). Anomalous states of knowledge as a basis for information retrieval. *Canadian journal of information science*, *5*(1), 133-143.

Boyce, B. R., Boyce, B. R., Meadow, C. T., Kraft, D. H., Kraft, D. H., & Meadow, C. T. (2017). *Text information retrieval systems*. Elsevier.

Chowdhury, G. G. (2004). Basic concepts of information retrieval systems. 2nd ed. London: Facet Publishing, 12.

Hearst, M. (2009). *Search user interfaces*. Cambridge university press.

Jones, K. S., & Willett, P. (Eds.). (1997). *Readings in information retrieval*. Morgan Kaufmann.

Kent, A. (1971). Information Analysis and Retrieval. New York: Becker and Hayes. *Inc., l97l*.

Unit–2

CATALOGING AND METADATA, SUBJECT ANALYSIS, AND REPRESENTATION

Complied by: Dr. Munazza Jabeen

Reviewed by: 1. Dr. Pervaiz Ahmad
2. Dr. Muhammad Arif
3. Muhammad Jawwad

CONTENTS

INTRODUCTION

For centuries libraries have been organizing reading materials on shelves for easy access. Researchers have found evidence of some form of cataloging activities for records held in the library of Alexandria in ancient Egypt around 300 BC. However, as far as modern cataloging and its objectives and principles are concerned, its history goes back just over two centuries. The first catalog code at the national level was the French Code of 1791. In Britain, cataloging rules were developed by Sir Anthony Panizzi for the British Museum library during the first half of the 19th century, and they were published in 1841.

However, systematic methods that have been widely adopted for the organization of library materials and their recording for use by readers came into being little more than a century ago. In 1876 Melville Dewey developed a systematic scheme of library classification, which became a unique tool for organizing library materials on the shelves, and in the same year Charles A. Cutter brought out Rules for a Dictionary Catalog, which enabled librarians to record systematically the library holdings in the form of catalog entries that could be consulted easily by the user community. Since then, several schemes of library classification and catalog codes have been developed to aid the process of organizing library materials systematically.

Details of bibliographic classification appear in the following unit. This unit discusses the basic principles of cataloging and gives guidelines related to the cataloging of internet resources. This unit also discusses the concept of metadata with reference to various metadata standards.

OBJECTIVES

After reading this unit, you would be able to:

i. Know about the information and its basic principles of cataloguing with guidelines related to the cataloguing of internet resources.

ii. Learn essential metadata standards for internet resources, museum objects, government documents and archival records.

iii. Understand the primary criteria for selecting effective software for cataloguing.

iv. Identify essential elements of metadata for cataloguing resources.

v. Understand the vital metadata management practices

2.1 CATALOGING

Harrod's Librarian's Glossary defines a catalog as "A list of books, maps, and other items, arranged in some definite order". It records, describes, and indexes (usually completely) the resources of a collection, a library, or a group of libraries. A library catalog is said to be the key to a library's collection as each catalog entry, containing the bibliographic details of a particular document, informs the user about the holdings of the library. The art of preparing catalogs is cataloging. Systems thinking was introduced into the discipline of information organization in 1876 by Cutter who was the first to recognize the importance of stating formal objectives for a catalog.

Why Cataloging?

The following major objectives of a catalog have been identified in the literature:

- to enable a person to find a book by:
  - author
  - title
  - subject
- to show what the library has:
  - by a given author
  - on a given subject
  - in a given literature
- to assist in the choice of a book:
  - by edition
  - by character

Implications for Opacs

OPACs are the interfaces that help users communicate with the collection(s) of a library. Typically, OPACs allow users to search the library's catalogue and provide some other facilities, such as checking borrower records, reserving reading materials and supplying library news bulletins. Although OPACs were first used in the mid-1970s, it was only at the beginning of the next decade that a significant number of libraries switched from card catalogues to automated ones. However, those first catalogues were usually modules linked to the automated circulation system and had brief catalogue records and very limited functionality. New Generation OPACs.

Cataloging of Internet Resources

Internet resources have some specific characteristics that call for some special rules for cataloguing. These characteristics include the rapidly growing number of resources; their availability through various agencies and the ways they are made

available — access to a large collection in one go (through licensing agreement, for example) instead of a gradual growth in number; and the need for regular management and maintenance due to their changing nature (including changes in location and terms of availability).

2.2 FUNCTIONAL REQUIREMENTS OF BIBLIOGRAPHIC RECORDS

Users of a library catalogue perform several tasks with the catalogue, for example:

- to find one or more items from the library's collection by conducting a search with one or more search keys
- to identify one or more items with some specified features
- to confirm that the retrieved items correspond to those looked for
- to select or choose one or more items in accordance with their specified content or format
- to obtain or acquire one or more selected items.

2.3 METADATA STANDARDS

Digital libraries and use of the internet have led users to be increasingly aware of the need for metadata for diverse categories of items available in digital form

Group	Element	Description
Content	Title	Name of the resource
	Subject	Topic describing the content of the resource
	Description	About the content of the resource
	Type	The nature or genre of the content of the resource
	Source	A reference to a resource from which the present resource is derived
	Relation	A reference to a related resource
	Coverage	The extent or scope of the content of the resource
Intellectual property	Creator	Who is primarily responsible for creating the content of the resource
	Publisher	Who is responsible for making the resource available
	Contributor	Who makes contributions to the content of the resource
	Rights	Information about rights held in and over the resource
Instantiation	Date	Date associated with the resource
	Format	The physical or digital manifestation of the resource
	Identifier	A unique reference to the resource within a given context
	Language	A language of the intellectual content of the resource

Table 2.1 Dublin Core data elements

Subject experts have developed, or are engaged in developing, various metadata formats for materials in specific domains, or for materials of specific kinds and formats, for example, metadata for internet resources, museum objects, government documents and archival records. There are two distinct schools of thought that influence the development of metadata standards:

- the minimalists camp whose point of view reflects a strong commitment to the notion of the simplicity of metadata for creation by authors and for the use of the metadata by tools
- the structuralists camp whose members emphasize the greater flexibility of a formal means of extending or qualifying elements so that they can be made more useful for the needs of a particular community.

2.4 DUBLIN CORE

The Dublin Core Metadata Initiative began in 1995 with a workshop that brought together librarians, digital library researchers, content experts, and text-markup experts to develop discovery standards for electronic resources. The first meeting took place at Dublin, Ohio, and gave rise to a metadata format called the Dublin Core. Table 2.1 shows the 15 Dublin Core data elements"" (also published as ISO Standard 15836-2003). The elements fall into three main groups, which roughly indicate the class or scope of information stored in them:

- elements related mainly to the content of the resource
- elements related mainly to the resource when viewed as intellectual property rights related mainly to the instantiation of the resource.
- The Dublin Core Metadata Editor is a service that retrieves a given web page and automatically generates some parts of the Dublin Core metadata for the given resource suitable for embedding at the head (in the <head>...</head> section) of the page.

In addition to instantly generating some of the Dublin Core tags for a given web page, the DC Dot service also provides an editor for the users to edit tags or add or edit contents, which can then be resubmitted to create metadata. Table 2.1 shows the Dublin Core metadata for a sample web page. Dublin Core standard has the following characteristics:

- The core set can be extended with further elements, as necessary, for a particular domain.
- All elements are optional.
- All elements are repeatable.
- Any element can be modified by a qualifier.

2.5 METADATA MANAGEMENT

Metadata can be embedded within the information resources, as is the case with web resources, or held separately in a database. Although metadata plays a big role in the resource discovery process, end-users don't see, and in most cases don't need to see, metadata for information resources that they are looking for. Metadata is mostly seen and used by information professionals who are involved in the organization and processing of information and is used by computer programs for several purposes such as resource identification, sharing, and interoperability.

2.6 DISCUSSION

Although cataloging remains highly relevant in the modem information retrieval environment, many parts of the catalog codes specifying rules for several activities have become redundant in the context of OPACs. AACR2 was not specifically designed to handle internet resources, and additional measures are required to catalog them. Nevertheless, AACR2 has played a key role in standardizing information retrieval activities (especially for OPACs) throughout the world for over four decades.

2.7 ACTIVITIES

1) Prepare a chart of important elements of the Dublin Core metadata standard to describe resources of specific kinds and formats in your project.
2) Visit a university library cataloging section. Conduct an interview with a cataloging librarian about how he/she narrates the importance of cataloging.
3) Transcribe the discussion and present it in your class for feedback from your class tutor.
4) Visit an automated university library and do practice on the cataloging interface to record and find some internet resources, observe if it appeals to you? Make your own experience instead of circulation employees and determine the ease of use, functionality, and appearance. Evaluate the cataloging software using the checklist provided in Section 2.3.

2.8 SELF-ASSESSMENT QUESTION

1. How does preparing a chart of metadata elements of your internet resources play an important role in selecting appropriate standards?

2. How meeting with cataloging librarians can help you to fulfill the cataloging requirements in your project?

3. How can metadata creation help mitigate the potential problems to the resource discovery process? What features of the discovery process should be included in the evaluation of good standards?

2.9 REFERENCES

Salton, G. (1989). Automatic text processing: The transformation, analysis, and retrieval of. *Reading: Addison-Wesley*, *169*.

Lancaster, F. W. (1968). Information retrieval systems; characteristics, testing, and evaluation.Ranganathan, S. R. and Gopinath, M. A., Colon Classification, 7th edn, Bangalore,

Foskett, A. C. (1996). *The subject approach to information*. Facet Publishing.

Mcilwaine, I., & Buxton, A. B. (2000). *The Universal Decimal Classification: a guide to its use*. The Hague: UDC consortium.

Chowdhury, G. G., & Chowdhury, S. (1999). Digital library research: major issues and trends. *Journal of documentation*.

Unit–3

AUTOMATIC INDEXING, FILE ORGANIZATION AND VOCABULARY CONTROL, ABSTRACTS, AND ABSTRACTING

Complied by: **Dr. Munazza Jabeen**

Reviewed by: **1. Dr. Pervaiz Ahmad**
2. Dr. Muhammad Arif
3. Muhammad Jawwad

CONTENTS

INTRODUCTION

The automatic indexing is defined as ‘when the assignment of the content identifiers is carried out with the aid of modem computing equipment the operation becomes automatic indexing’. Borko and Bernier suggest that the subject of a document can be derived by a mechanical analysis of the words in a document and by their arrangement in a text. In fact, all attempts at automatic indexing depend in some way or other on the original document texts, or document surrogates. The words occurring in each document are listed and certain statistical measurements are made, such as word frequency calculation, total collection frequency, or frequency distribution across the documents of the collection.

Vocabulary control is one of the most important components of an information retrieval system. As we have noted from its simple model given in unit I, an information retrieval system tries to match user queries with the stored documents (document surrogates) and retrieves those that match. To match the contents of the user requirements (the search terms) with the contents of the stored documents (the document records), one must follow a vocabulary that is common to both. In other words, user requirements need to be translated and put to the retrieval systems in the same language as was used to express the contents of the document records. This leads us to the concept of using a standard or controlled vocabulary in an information retrieval environment.

An abstract is different from an extract, an annotation, or a summary. An extract is an abbreviated version of a document created by drawing sentences from the document itself, whereas an abstract, although it may include words that appear in a document, is a piece of the text created by the abstractor rather than a direct quotation from the author. An annotation is a note added to a title or other bibliographic element of a document by way of comment or explanation, and a summary is a restatement within a document (usually at the end) of the document’s salient findings and conclusions.

OBJECTIVES

After reading this unit, you would be able to:

i. Learn the manual indexing process to automatic indexing in the era of Big Data and Open Data.

ii. Make a difference between direct & sequential access to peripheral devices required for an automated library system.

iii. Understand the essential criteria for binary search and binary search trees.

iv. Learn the role of vocabulary control for an effective information retrieval system, and what are the practical vocabulary control tools?

v. Understand the essential criteria for a subject heading list to represent the subject content of an information resource

vi. Differentiate abstract by writer and abstract by purpose to convey the document's salient findings and conclusion

3.1 THE PROCESS OF INDEXING

Before going into much detail of the process, we should try first to understand the advantages of automatic indexing. Salton mentions the following:

- level of consistency in indexing can be maintained
- index entries can be produced at a lower cost in the long run
- indexing time can be reduced, and
- better retrieval effectiveness can be achieved.

Harter points out that automatic analysis by means of word frequency analysis can be viewed as a two-tiered problem." In the first stage, the problem relates to the identification of a technical vocabulary characteristic of a given subject field. Once the vocabulary or index terms have been chosen, the second problem arises, which relates to the representation of the document with the help of keywords.

3.2 INDEX FILE ORGANIZATION

The elementary units of a text retrieval system are document records. Each document record comprises several fields and subfields, each one of which contains a particular unit of information — author's name, publisher's name, title, the keyword(s), class number, ISBN, and so on. The document record may also contain the abstract or full text of the document concerned. A text retrieval system is designed to provide fast access to the records through any of the sought keys or access points. This means that there should be a mechanism for fast access to the document records. What should the basic mechanism be for accessing the document records through some key values — by chosen keyword(s), or by author's name, say? To answer this question, we should first understand how document records are physically stored in the computer.

3.3 INVERTED FILE

In an inverted file system of text retrieval, each database consists of two files. One is the text file, which contains what we would expect to find, that is the document records in their normal form — the form in which they are entered into the database. The other is the inverted file, which contains all the index terms, drawn automatically from the document records according to the indexing technique adopted for the purpose. Each index term in the inverted file is associated with a pointer that shows the record number in which the index term occurs.

(a) Document records

Document No. 1

Author: Cunningham, ill.

Title: File structure and design Publisher: Chartwell-Bratt Year: 1985

Keywords: File structure; File organization

Document No. 2 Author: Tharp, A.

Title: File organization and processing Publisher J oh n WileyYearr: 1988
Keywords: File structure; File organization

Document No. 3 Author: Ford, N.

Title: Expert systems and artificial intelligence Publisher. Library Association
Year: 1991
Keywords: Expert systems; Artificial intelligence; Knowledge-based systems

Document No. 4

Author: C harnHarnic; McDe rmott, D. Title. Introduction to artificial intelligence PPublisher Ad diso n-Wesley
Year: 1985
Keywords: Artificial intelligence; Expert systems

3.4 SAMPLE DOCUMENT RECORDS AND SAMPLE INDEX

Figure 3.1 shows the essence of an inverted file approach. However, an inverted file may contain a lot of other information along with each entry, such as the number of occurrences of the term in each record; or position information such as the field in which the term/phrase occurs, or where the term/phrase occurs in each sentence/paragraph.

The field tag is used to denote the field where the given term/phrase occurs. This information is used in field-specific searches. Similarly, the position information is used for proximity or adjacency searching. Other types of information may also be stored along with each entry, and each such item of information facilitates a particular type of search. Nevertheless, the more such information is added to each entry, the bulkier the inverted file becomes, therefore taking up more storage space and needing more processing time. In this example, a user looking for the phrase ‘expert systems will retrieve two records, document numbers 3 and 4 from the database, while another user looking for a book written by ‘Tharp, A.’ will retrieve book number 2. A complex query with search terms combined with Boolean operators will follow the same path. For example, a user with a query ‘expert systems OR file organization’ will retrieve all four document records, while the query ‘artificial intelligence AND knowledge-based systems will retrieve document record number 3. In the first example, as the search terms are joined by the logical operator ‘OR’, the system will consult the inverted file for each term and will then merge the document numbers retrieved in each case; while in the second, because the terms are joined by the logical operator ‘AND’, the retrieved

document numbers for both terms will be matched to locate the common document numbers, i.e. the ones where both terms are present.

Figure 3.1 shows that each term may occur in a few documents (for example, Term l occurs in Docl and Doc5), and in each case we need to store information on the number of occurrences (O), field of occurrence (F), position information (P), and so on. Thus, for many terms, the index file may be quite large and complex. To avoid this, in the inverted file organization, information about index terms is stored in two different files. Let us take a simple example: suppose, we have a file of 10,000 documents for which there are 1000 index terms. Two different files can be created to store information about the index terms. The first file may be quite short, containing only 1000 entries, each entry having only three fields: where field 1 contains the index term, field 2 contains the frequency of occurrence (this information is used for several purposes in a search), and field 3 contains the address of the block containing the addresses of documents whose document profiles include the descriptor from field 1. Such an index file can easily fit into the primary storage where a fast search for a required search term can be performed. The second file consists of several blocks were.

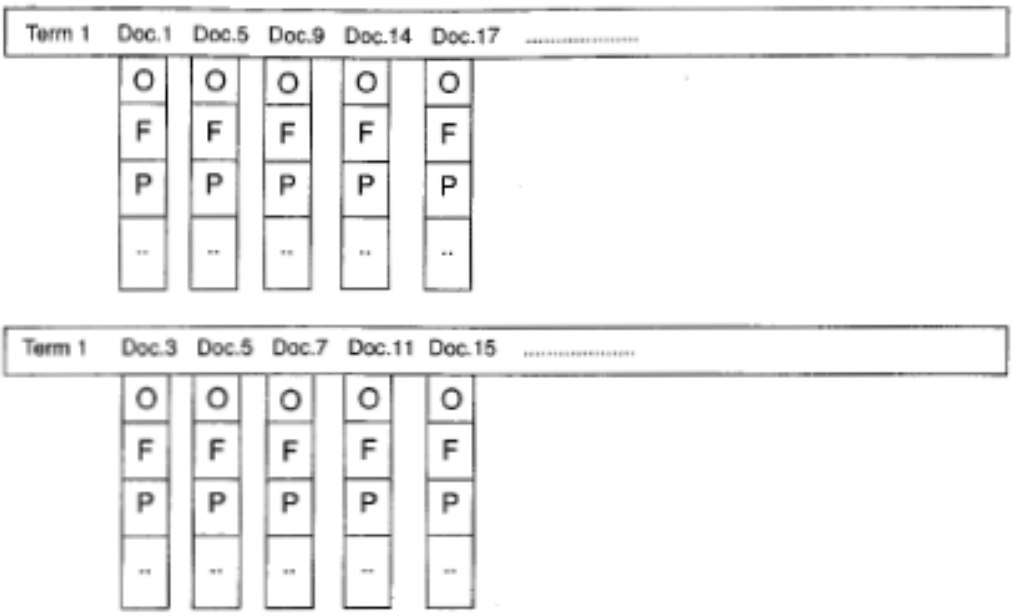

Figure 3.1 A simple view of an entry in an inverted file

Each block contains the addresses and other associated information of those documents where the given search term occurs. The second file may be quite large

because each index term may have occurred in a number of records, and therefore, some blocks may contain several lists of addresses. This is handled by linked lists and pointers. Figure 3.2 shows such an index file.

Figure 3.2 shows that for Term49, we need to store only the address for the first record where it occurs, which in this case is 105. Thus, a pointer from the first file points to an address block where the document and the associated information is stored. Here, after the information about the first document, there is another pointer leading to address block 612, another pointing to address block 911, and finally a null pointer (/\ ) indicating that it is the end of the list. Thus, we need only one address for each descriptor in the index file. This is the address of the first block containing the address of document indexed by the given descriptor, which may lead to the subsequent address blocks each containing the document number and other associated information.

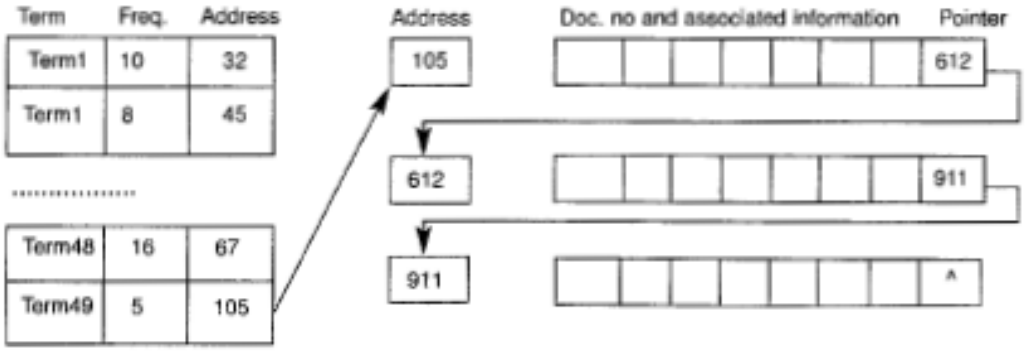

Figure 3.2: Index file and linked list

3.5 ACCESS TO INVERTED FILES

The user may pose a single key query or a multiple key query. In the former case, the value of a single search key (say the name of the author) is used as the retrieval criterion, whereas in a multiple key search a number of search keys (say the name of the author, subject name, date of publication, and so on, as in the query ‘papers written by Salton on information retrieval systems between 1980 and 1990’). For single key searches, the whole file can be maintained in an order according to the value of the given single set of keys. In a telephone directory, for example, users search through the names of subscribers and therefore the names of subscribers are arranged in alphabetical order. File access in multi-key searches is complicated by the fact that it is not possible to order the file simultaneously in accordance with the values of the different search keys. For example, a users’ file in a library can be

arranged according to the name of the user, occupation or specialization, address or department, and so on, and in each case the resulting arrangement of the records within one field will be different from the other.

In the case of a multi-key search, a principal key is to be identified and the file can be ordered in accordance with the values of that key. When the principal key is used as part of a search statement, the subsection of the file corresponding to the given principal key value can then be isolated and subjected to a separate search based on the values of any secondary keys also included in the search query. A catalogue of a library can be considered as a multi-key file, where the keys are the author, title, publisher, subject and so on. In such a file, the principal key is usually the author: the file is ordered in accordance with the name (surname) of the authors. From each record in the main file there may be a number of pointers giving access to secondary keys, such as publisher and title. A simple file of authors and publishers can be ordered according to the author's name as the principal key, with a sparse index giving access to a chain of pointers for each publisher name. Documents published by a given publisher can be found by following the pointer chain. Pointer chains can be provided for all secondary keys in addition to the primary keys attached to the records; each given record can be traced through the pointer chain for any of the keys. This type of record organization is known as a multi-list.

3.6 SEQUENTIAL ACCESS

As mentioned earlier, the easiest approach to adopt for consulting an inverted file is the sequential access method that is, looking for each index term one after another from the beginning until either the sought term is found, or a key with a higher key value is reached, or the end of the index file is reached.

However, the time taken for a sequential search depends on two factors:

1. The key value of the sought key, because the placement of the key will be based on its key value, and
2. The length of the index file.

Sedgewick that a sequential search has two major properties:

3. R Sequential searches in an unordered file use N + 1 comparisons (N is the number of records in the file) for an unsuccessful search and an average of N/2 comparisons for a successful search.
4. Sequential searches in an ordered file use about N/2 comparisons for both successful and unsuccessful searches.

3.7 ALPHABETIC CHAIN

One way to reduce the number of search probes in an ordered sequential file organization is to use an alphabetic chain. What is an alphabetic chain? Let's take a simple example. What do we do when we look for a term in a dictionary? Let's suppose that we are looking for the term 'psychology'. We don't start from the beginning of the dictionary but rather from the letter 'p', thereby skipping the other letters. Within 'p' we skip terms beginning 'pa', 'pb', and so forth, and start with the words beginning 'ps', then search for the word sequentially. In the same way, when we use an index we skip some part of the index file so as to reduce the number of words to be searched.

3.8 BINARY SEARCH

We have already seen that for a large index file the sequential search is quite a time-consuming process. An improvement would be a reduction in the number of probes required in conducting a search. If we have an ordered file of index terms, we can reduce the number of probes needed to retrieve an index term by applying a binary search technique. As the term 'binary' suggests, the basic mechanism followed is to divide the search file automatically into two parts at successive levels and to thereby reduce the number of search probes.

3.9 BINARY SEARCH TREE

One problem with the binary search technique is that at each stage the system has to determine the middle term and only then can the search continue. This could be avoided if there was a mechanism by which all the terms in the index file could be organized according to their key values in such a way that at each stage only two options would be available, and the system could proceed by making only one comparison. In fact, the binary search is most conducted by using a tree structure of records. A tree structure consists of nodes or vertices containing node information, together with pointers giving access to additional nodes of the tree. The defining property of a tree is that every node is pointed to by only one other node, called the parent. A tree organization supports operations such as searching for a record, inserting new records and deleting records. Tharp defines a binary search tree as a binary tree in which the nodes are used both to store information and to provide direction to other nodes.21 In binary search trees information is organized in such a manner that all the key values smaller than that of the current node are stored in its left tree and those that are larger are in its right tree.

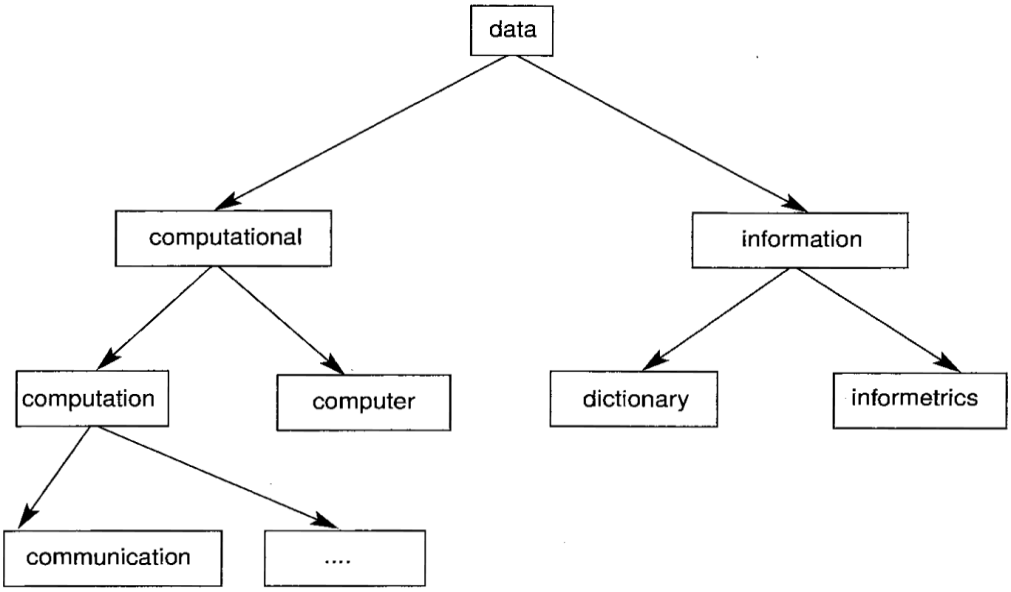

Figure 3.3 Simple binary search tree

In fact, the maximum number of key comparisons needed to conduct a binary tree search is equal to the longest path from the root to a leaf of the tree.

The insertion of a new term other than in a leaf node (that is a blank space suitable for an insertion) and deletion of a term from the tree sometimes requires a major reshuffle of the tree. In fact, node deletion in a binary tree is more complicated than node searching or node addition because the proper tree structure must be preserved when a node is deleted. To insert a key in a binary tree, first an unsuccessful search is conducted and then the node is inserted at the empty node where the search had terminated. Deletion of a node in a binary tree is quite a cumbersome job. Salton suggests the following steps for the deletion of a node from a binary tree:

1. If the node to be deleted is the leaf of the tree, then the corresponding node is simply deleted.
2. If the node to be deleted has only one child, that is either the left or the right node of this is empty, then the node to be deleted may be replaced by the only available child node.
3. If the deleted node has two children, then the deleted node information is replaced by the node with the smallest key value in the right subtree.

3.10 BALANCED TREE

A balanced tree (B-tree), in contrast with a binary tree, is a multi-way search tree. A binary tree has a branching factor of two, whereas a balanced tree does not have a theoretical limit to branching factors. A binary tree grows downwards as new terms are added, whereas a B-tree grows upwards with an increase in size.

3.11 CONTROLLED VS NATURAL INDEXING

Controlled indexing languages are those in which both the terms that are used to represent subjects and the process whereby terms are assigned to documents are controlled or executed by a person. Normally there is a list of terms, a subject headings list, or a thesaurus that acts as the authority list in identifying terms that may be assigned to documents and indexing involves the assignation of terms from this list to specific documents. The searcher is expected to consult the same controlled list during the formulation of a search strategy. In natural language indexing, any term that appears in the title, abstract, or text of a documented record may be an index term. There is no mechanism to control the use of terms for such indexing. Similarly, the searcher is not expected to use any controlled list of terms. Svenonius divides the debate concerning natural and controlled vocabulary into three eras, which Rowley modified into four, as shown in Table 3.1.

Table 3.1 The four eras of debate on controlled vs. natural language indexing

No.	Eras of debate
1	Controlled vocabulary
2	Comparisons of natural and controlled language: major experimental studies noted that natural language can perform as well as controlled vocabulary, but other factors, such as the number of access points, are also significant.
3	Many case studies of limited generalizability. Searching online databases was considered. It was noted that the best performance can be achieved by a combination of controlled and natural language; the number of access points was reaffirmed to have a significant effect; full-text and bibliographic databases were noted to have produced different results.
4	New advances in user-based systems including OPACs. The value of controlled vocabulary in the context of user-friendly interfaces and the development of knowledge bases were noted.

Aitcheson and Gilchrist provide a comparison of controlled and natural language indexing, which is shown in Table 3.1. Rowley mentions that despite much debate extending over more than a century, together with a range of research projects, information scientists have failed to resolve the issue concerning the relative merits

and demerits of controlled and natural language. However, practice and tested research have suggested that controlled language and natural language should be used in conjunction with one another.

3.12 VOCABULARY CONTROL TOOLS

As the name suggests, these are the tools used to control the vocabulary of indexing and retrieval. These are natural language tools, meaning that these tools contain natural language terms that can be used for indexing and retrieval purposes. What an indexer and an index user need is a set of guidelines for the proper selection of terms. Syndetic structures are devices that provide these guidelines by showing the relationships among terms or concepts, and they fall into two major categories: classification schemes, subject heading lists and thesauri. Classification schemes, being tools for organizing knowledge, could be of great help for vocabulary control but the main body of classification schemes are organized in an artificial language, whereas for vocabulary control we need natural language representation. Indexes to classification schemes could serve the role of vocabulary control but here terms appear alphabetically and thus the logical (semantic) organization of knowledge is not available. Some attempts have been made to (coordinate and collateral) terms. However, this distinction has gradually faded, and the latest Library of Congress subject headings list indicates the terms' features as shown in normal thesauri.

3.13 SUBJECT HEADINGS LISTS

A subject heading list is an alphabetical list of terms and phrases, with appropriate cross-references and notes, which can be used as a source of headings in order to represent the subject content of an information resource. Although it is primarily arranged alphabetically by term, under each term or phrase we can find a list of other terms or phrases that are semantically related to the term or phrase. Subject heading lists were designed to complement bibliographic classification in the sense that although a bibliographic classification scheme helps us to assign a class number (built of notations) to an information resource that represents its subject content, a subject heading list allows us to assign an appropriate heading, as a term or a phrase, to an information resource that represents its subject content. A list of subject headings, or a subject index as it is often called, can be used to search or browse a collection of information resources. Subject heading lists help us produce a pre-coordinated index of a collection.

Library of Congress Subject Headings (LCSH) is an example of a subject heading list; it is used widely as a controlled vocabulary for catalogs and bibliographies. It was originally designed as a controlled vocabulary for representing the subject and

form of the books and serials in the Library of Congress collection, with the objective of providing subject access points to the bibliographic records contained in the library's catalogues. It is now most widely used for assigning subject headings to bibliographic information resources.

3.14 TYPES OF ABSTRACT

Abstracts may differ according to their writer, purpose and style. Guinchat and Menou suggest that the various types of abstracts can be distinguished by:

- Their length, which normally ranges from a few dozen to several hundred words, and is occasionally over a thousand
- The amount of detail, which can differ significantly; certain abstracts (known as indicative abstracts) simply provide a brief summary, whereas others (known as informative abstracts) include a varying number of points that are likely to interest the user
- The inclusion of judgments or critical analysis, which may amount to some form of evaluation of the document
- Whether the indexer deals with the whole document or only with aspects that are likely to interest the user (known as slanted abstracts)
- Whether the author of the abstract is the author of the original document oz. Some other person, and
- The language used, which may be a natural language or a more formalized (artificial) language.
- Different criteria have been used by other information scientists to categorize the different kinds of abstracts, as discussed in the following sections.

Abstract by Writer

Abstracts may be written by authors, by subject experts or by professional abstractors. Thus, we may categorize them as: author-prepared abstracts, expert-prepared abstracts and professional-prepared abstracts.

Articles in professional journals are usually accompanied by author-prepared abstracts. While these abstracts, being written by the authors of the articles themselves, should contain the most sought-after information, they usually lack a professional style. Expert abstractors are usually the choice of abstracting journals. These persons are trained in abstracting and are also expert in the subject field. Thus, their abstracts should be accurate, comprehensive, lucid and terse. These abstracts are usually very promptly delivered and are well written, although they

may sometimes be expensive. Professional abstractors abstract for a living, and may be employed to handle work in more than one language.

Abstracts by Purpose

Abstracts are written with certain purposes in mind, and therefore there may be different sorts of abstracts to serve different purposes. Borko and Bemier have identified four different types: the indicative abstract, informative abstract, critical abstract and special purpose abstract.

An indicative abstract simply indicates what the parent document is all about. They are also called descriptive abstracts, because they usually describe what can be found in the original document. Indicative abstracts may contain information on purpose, scope or methodology, but not on results, conclusions or recommendations.

Thus, an indicative abstract is unlikely to serve as a substitute for the original document.

An informative abstract is intended to provide readers with quantitative and qualitative information as presented in the parent document. Informative abstracts may include information on purpose, scope, and methods, as well as the results and findings. Informative abstracts are often longer than other types of abstracts and are difficult to write, but they often save the user the time necessary to consult the original work.

A critical abstract contains critical comments or review by the abstractor. For indicative and informative abstracts, abstractors usually function as impartial reporters, whereas in a critical abstract, the abstractor deliberately includes his/her own opinion and interpretations. Preparation of critical abstracts requires subject expertise and is a time-consuming job.

Some abstracts may have been written to serve a special purpose or with a specific category of users in mind. Such abstracts are called special-purpose or slanted abstracts. Depending on the nature of the target user group, an abstractor may stress some part of the abstract (with more emphasis on in formativeness) at the expense of some other part(s) (leading to an indicative abstract for that part). Some abstracts may have a slant towards some part of the subject dealt with in the original document; these are particularly useful for mission-oriented works rather than in discipline-oriented works.

Lancasters and Borko and Berniee suggest that another category of abstract can be identified in this group, called the modular abstract. Here an abstractor is expected to prepare different kinds of abstracts — indicative, informative, critical, and so on — any one of which may be used depending on the requirement of the abstracting agency. In fact, the abstractor writes various modules of abstract at the same time. Modular abstracts are intended as full content descriptions of current documents in five parts: a citation, an annotation, an indicative abstract, an informative abstract and a critical abstract. The prime purpose of modular abstracts is to eliminate the duplication and waste of intellectual effort in the independent abstracting of the same documents by several services, without any attempt to force 'standardized' abstracts on services whose requirements may vary considerably as to form and subject slant.

Qualities of Abstracts

An abstract must be brief and accurate, and it must be presented in a format designed to facilitate the skimming of a large number of abstracts in a search for relevant material. Guinchat and Menou suggest that an abstract should possess the following qualities; it should be:

- concise: whilst it should not be done at the expense of precision, however long the abstract is, care should be taken to avoid expressions or circumlocutions that can be replaced by single words
- precise: one should use expressions that are as exact and specific as possible without exceeding the abstract's requested length
- self-sufficient: the description of the document should be complete in itself and fully understandable without reference to any other document
- objective: there must not be any personal interpretation or value judgement on the part of the abstractor (obviously this does not apply to critical abstracts).
- Borko and Bernier give the following basic qualities of abstracts:
- Brevity. one of the essential characteristics of abstracts is their brevity: they are much shorter than the documents from which they are derived. Brevity saves the user's time, and it lowers the cost of producing abstracts. However, it must be remembered that, while redundancy is to be avoided, there should not be any loss of novelty when trying to achieve brevity.
- Accuracy: abstracts should be accurate, and errors avoided as far as is practicable. Errors may occur at many stages in the production of abstracts: in understanding the document's content and presentation, in the citation, and in typing, printing, and so on.
- Clarity: while an abstract should be brief and accurate, it must also be clearly written, avoiding all sorts of ambiguities.
- A good abstract should also have the following qualities: 9

- be a self-contained unit, a complete report in a miniature form; it should be intelligible without reference to the original document
- enable its users to (a) identify the basic contents of work quickly and accurately, (b) determine its relevance to their interests, and (c) decide whether or not to read the original document in its entirety
- be capable of being used as a secondary source of information R be impersonal
- not take a critical form (except for critical abstracts) R be as up to date as possible
- be able to be used as a retrieval aid in an automated information retrieval environment
- not repeat the information that is obvious from the title or that is well known to the user
- avoid redundancy and repetition
- be written in a clear and natural language and should avoid using abbreviations.
- Borko and Bemier comment that without surrogates, such as abstracts, searching through the accumulated literature would be an impossible task. In fact, there are a number of uses of abstracts, and that is why abstracting journals (in hard copy and/or on CD-ROM) have existed in almost all subject fields all over the world. Guinchat and Menou identify three major functions of abstracts.

3.14 ACTIVITIES

1) Identify which indexing method will best suit your documents of the collection through research, investigation, and a close examination of your search needs in terms of word frequency calculation, total collection frequency, or frequency distribution across the documents of the collection.
2) Consider your library has decided to enable a controlled vocabulary for representing the subject and form of the books and serials in the library collection. What kind of measures you will take to enable it with the objective of providing subject access points to the bibliographic records contained in the library's catalogs?
3) Consider yourself an editor of a professional journal? How will you see the pros and cons of author-prepared, expert-prepared, and professional-prepared abstracts? Which method would you prefer for abstracting, and why?

3.15 SELF-ASSESSMENT QUESTION

1. Why articles in professional journals are usually accompanied by author-prepared abstracts?

2. Why the indexes to classification scheme could not be an effective option to serve the role of vocabulary control?

3. Why expert abstractors are usually the choice of abstracting journals?

3.16 REFERENCES

Pao, M. L., *Concepts of Information Retrieval*, Englewood, CO, Libraries Unlimited,1989**.**

Atherton, P., Handbaok of Information Systems and Services, Paris, Unesco, 1977.

Guinchat, C. and Menou, M., General Introduction to the Techniques oflnformation and Documentation Work, Paris, Unesco, 1983.

Taylor, R., Question-negotiation and Information Seeking, *College & Research Libraries,* 29 (3), 1968, 178–94.

Xie, I., Interactive Information Retrieval in Digital Environments, Hershey, IGI Publishing, 2008.

Belkin, N. J., Oddy, R. N. and Brooks, H. M., ASK for Information Retrieval, Part-1, background and theory, *Journal of Documentation,* 38 (2), 1982, 61–71.

Unit–4

SEARCHING AND RETRIEVAL AND USERS OF INFORMATION RETRIEVAL AND USER-CENTERED MODELS OF INFORMATION RETRIEVAL

Complied by: **Dr. Munazza Jabeen**

Reviewed by: **1. Dr. Pervaiz Ahmad**
2. Dr. Muhammad Arif
3. Muhammad Jawwad

CONTENTS

INTRODUCTION

Users interact with an information retrieval system through an interface and several activities are performed: users' queries are received and interpreted, appropriate search statements are formulated, and the actual search (matching queries with the document profile or database) is conducted with a view to retrieving the required information. All these tasks can be performed manually, as used to be done in the earlier systems, or can be automated. The development of cheaper direct-access mass storage devices, magnetic disks and drums, the associated software, and advancements in electronic communication systems brought about the possibility of more dynamic searching via online methods. The concept of online searching has occupied a large and significant area in the study and research of modern information retrieval. However, a user often faces difficulties in approaching an online search system, especially in formulating an appropriate search statement. The cost of searching a database, whether in-house or external, can be reduced significantly if an appropriate strategy for searching is followed. The search strategy helps the user select the optimum path for searching a file or a database. This involves several measures that are to be taken before and during a search. This unit discusses the basic concepts of the search strategy and describes the actual searching process in the context of information retrieval systems. Features of online searching are discussed later.

The user is the focal point of all information retrieval systems because the sole objective of any information storage and retrieval system is to transfer information from the source (the database) to the user. The characteristics and specific needs of users determine the nature of the information to be collected by the system, the nature and level of analysis to be made to store the information, and the nature of the user interface to be designed so that users can interact with the system easily to search and retrieve the required information. Thus, an understanding of the nature and number of users, their activities vis-a-vis information requirements, information-seeking behavior, and so forth, will help an information manager develop an appropriate information retrieval system.

OBJECTIVES

After reading this unit, you would be able to:

i. Familiar with different information search strategies and how the information search can be broadly divided into different categories.

ii. Learn to get the correct understanding of the users' precise needs through pre-search interviews.

iii. Understand the essential criteria for the searching process.

iv. Identify influential information retrieval models and ways to approach them for designing and implementing retrieval systems?

v. Learn fundamental needs of user information for retrieval systems.

vi. Understand the basic information required to support community development planning.

4.1 THE SEARCH STRATEGY AND ITS PREREQUISITES

Information search can be broadly divided into the following major categories:

1. *Known item search:* The searcher knows about the existence of a certain piece of information and wants to find it in a specific collection. In the context of an online public access catalogue, this can be a search for a book written by a specific author, with a specific title, and so on. In the context of the web, this can be the search for the web page of a specific department or faculty. These searches are usually not complicated and can be accomplished relatively easily and quickly.
2. *Search for specific information or a fact**:*** Users may often search for a certain piece of information such as who is the current Secretary-General of the United Nations, or what the population of India is. Such information may be obtained from a variety of sources, typically from reference books or reference databases and websites, and often the user can find the required information either directly as an answer or through a reference to a text that contains the information.
3. *Search for information related to a problem or issue:* This is the most difficult type of information search, for various reasons: the user may not know exactly what they want; the information may be available from a variety of information channels and sources; the information may need to be gathered and aggregated or synthesized, or the user's information need may change on receipt of some information (the user may find that what they were looking for in the first instance was not quite what they wanted). The nature of the user's problem, knowledge, or subject background may have a significant influence on the search process and relevance judgment. These searches are often demanding in time and expertise required.
4. *Exploratory search:* This type of search may be rather undirected apart from the fact that the searcher wants to know about the content of a database or a website. This kind of search may often help the user find some useful information that may not have been asked for specifically; and this may also lead to an accidental discovery of information, called serendipity.
5. *Search to keep up to date in a specific field:* Specialist users often want to stay up to date in their field, and that's why they regularly search or scan various journals and databases. Traditionally this kind of search service has been offered through what is known as current awareness services (CAS), or selective dissemination of information (SDI) services. Nowadays, several special programs are available that automatically search for users in chosen subjects and topics, specific databases, and websites.

The search strategy may be defined as a plan for conducting a search for information, and therefore should include a search objective and a plan of operation. It encompasses several steps and levels of work in information retrieval. Meadow and Cochrane mention that the search strategy includes at least three decision points that a searcher has to reach. There are many issues that need to be considered while formulating an appropriate search statement:
The concepts or facets to be searched and their order
The term(s) that appropriately represent(s) the search concept
The feature(s) of the retrieval system concerned

4.2 THE PRE-SEARCH INTERVIEWS

The results of a search depend heavily on the correct understanding of the users' precise needs. This understanding can be developed through a pre-search interview. This is crucial if the client is not to be present during the actual search.
A pre-search interview is a conversation that takes place between a user and member of the information staff on the information requirement of the user. Somerville lists the following skills that the successful pre-search interviewer should possess:

- Ability to conduct personal communication well
- Conceptual skills
- Analytical skills
- Knowledge of file organization
- Understanding of indexing policy and vocabulary control
- Subject knowledge.

It may be noted that the concept of a pre-search interview presupposes the existence of a search intermediary. This concept was developed to get a better understanding of the search requirements of a user in an online search environment. As discussed later in this book, the concept of a pre-search interview, although very important in a mediated search process, has very little relevance in the context of information retrieval from the world wide web and digital libraries, since these systems are designed to be used by end-users without any direct involvement of human intermediaries. Nevertheless, an understanding of the process involved in a search interview process may be useful for designers of information retrieval systems in the web and digital library environment.

4.3 THE SEARCH PROCESSES

A database may comprise controlled or uncontrolled vocabulary. Cleverdon mentions that a user searching a database that has controlled index languages must do the following:

1. Decide the words that might be used by the authors of the relevant documents.
2. Decide which database(s) is/are to be searched.
3. A Use the thesaurus of the chosen database to translate the query terms in the appropriate way.
4. Guess which of the chosen terms (or concepts) might have been used by the database indexer.
5. Co-ordinate the terms (often using Boolean operators) to formulate the search statement.
6. Input the search statement.
7. Repeat steps 5 and 6 until a desirable output is obtained or the search fails altogether.
8. Identify the actual relevant items from among those retrieved.
   One major task in the searching process relates to the coordination of terms (step 5 above) to formulate the actual search statement. The result of the search depends largely on how adequately the search terms are combined. Boolean search techniques have been used widely since the beginning of mechanized information retrieval.

4.4 RETRIEVAL MODELS

Since the early days of information retrieval research, efforts have been made to develop formal theory-based approaches to model various aspects of information retrieval systems. An advantage of adopting a model-based approach is that each model, although different from another, is based on a set of principles and assumptions, and, while theory drives experiment in suggesting new ways and means of carrying out tests, at the same time experiment drives theory by either justifying or improving the model.

Information retrieval models can broadly be classified into two groups: user-*centred/cognitive* models, and *system-centred* models. Cognitive models take a holistic view of information retrieval, by considering not only the retrieval mechanisms used in matching the queries with stored information but also the following:

- The ways in which the user's information need can be formulated as a query.
- The human-computer interactions that take place during the search process.
- The social and cognitive environments in which the process takes place.
- How the information is used by the user to meet a specific information need.

These models are very different in their nature, and in their disciplinary bases, when compared with the system-centered models, which provide an explicit statement of the workings of the information search and retrieval mechanism. User-centered information retrieval models. This unit provides an overview of classical system-centered) information retrieval models. These include: the Boolean search model,

which compares Boolean query statements with the term set used to represent document contents; the probabilistic retrieval model, which is based on the computation of relevance probabilities for the documents of a collection; and the vector processing model, which represents both documents and queries by term sets and compares global similarities between queries and documents. Several models have been developed based on the classical information retrieval models, not all of which have been discussed in this book, but appropriate references have been provided for interested readers. While classical retrieval models are based on logical and mathematical principles, some alternative models of information retrieval have also been developed over the past few years. Two prominent alternative types of retrieval model.

4.4.1 BOOLEAN SEARCH MODEL

George Boole (1815-64) devised a system of symbolic logic in which he used three operators: +, x, and —, to combine statements in symbolic form. John Venn later expressed Boolean logic relationships through what are known as Venn diagrams. The three operators of Boolean logic are the logical sum (+), logical product (x) and logical difference (—). Information retrieval systems allow the users to express their queries by using these operators.

4.4.2 LIMITATIONS OF BOOLEAN SEARCHING

Despite its popularity, the Boolean search model has certain limitations. The first relates to the formulation of search statements. It has been noted that users are not always able to formulate an exact search statement by the combination of AND, OR and NOT operators, especially when several query terms are involved. In such cases either the search statement becomes too narrow or too broad. Boolean searching, therefore, often calls for a trained intermediary.

4.4.3 PROBABILISTIC RETRIEVAL MODEL

Probability theory has been used as a principal means for modelling the retrieval process in mathematical terms. In conventional retrieval situations a document is retrieved whenever the keyword set attached to the document set appears similar in some sense to the query keywords. In this case, the document is considered relevant to the query. Since the relevance of a document with respect to a query is a matter of degree, it can be postulated that when the document and query vectors are sufficiently similar, the corresponding probability of relevance is large enough to make it reasonable to retrieve the document in response to the query.

The basic underlying tenet of the probabilistic approach to retrieval is that, for optimal performance, documents should be ranked in order of decreasing probability of relevance or usefulness to the user. Probabilistic approaches,

therefore, attempt to estimate or calculate, in some way, the probability that a document will be relevant for a particular user. Several models based on probabilistic approaches have been advocated; here we shall briefly look into three such models.

4.4.4 THE VECTOR PROCESSING MODEL

The vector processing model assumes that an available term set, called term vectors, is used for both the stored records and information requests. Collectively the terms assigned to a given text are used to represent text content.

Consider a collection of documents in which each document is characterized by one or more index terms. Thus, the documents are the objects in the collection each of which is represented by a number of properties (here index terms). The similarity between two objects is normally computed as a function of the number of properties that are assigned to both objects; in addition, the number of properties that is jointly absent from both the objects may also be taken into account. Substantially similar methods can be used for determining collection structure (by comparing pairs of text vectors with each other and identifying text pairs found to be sufficiently similar), and for retrieving information (by comparing the query vectors with the vectors representing the stored items and retrieving items that are found to be similar.

4.4.5 BEST MATCH SEARCHING AND RELEVANCE FEEDBACK MODEL

Best match searching is designed to produce ranked output. It therefore requires a method to measure the relative importance of the retrieved items, which again requires some method of weighting the search terms. A similarity measure comprises two major components: a term weighting scheme that reflects the importance of a term by allocating numerical values to each index term in a query or document; and a similarity coefficient which uses these weights to calculate the similarity between a document and a query. Of these, the term weighting scheme is the most important, and, as Sparck Jones and Willett comment, it is the single most important factor in determining the effectiveness of an information retrieval system.

The best match search matches a set of query words against the set of words corresponding to each item in the database, calculates a measure of similarity between the query and the item, and then sorts the retrieved items in order of decreasing similarity with the query. Cleverdon suggested the so-called coordination level to measure the relative importance of the retrieved items. According to this theory, the documents that have the most terms in common with the query will be on the top of the ranked list. A best match search can be

implemented very efficiently using an inverted file searching technique. The user in a best match search environment can put the query in simple natural language, in the form of a sentence, say. The terms representing the query, or a document are then identified, and measures are taken to overcome the variations due to spelling, synonyms, antonyms, and so forth. There is thus the need for a conflation algorithm, a computational procedure that reduces the variants of a word to a single form for retrieval purposes. The most common automatic conflation procedure uses a stemming algorithm, which reduces all the words with the same route to a single form by stripping the root of its derivational and inflectional affixes in most cases only suffixes are stripped.

4.5 USERS AND THEIR NATURE

The concept of the user is by no means clear. The type of information users in fact depends on the nature of the information; users may be limited by the organization they work for, the nature of their work or profession, their age, sex, or other social groups, and so on. Several criteria may be used to identify and categorize users. For example, user categories may be identified by the nature of the libraries they use:

- For an academic library, primary users are students, teachers, researchers, and to some extent administrators.
- For special or research libraries, primary users can be determined by the nature of their work or profession, or by attachment to the parent organization; they may be categorized as researchers, planners and policymakers, managers, engineers, doctors, scientists, agriculturists, and so on.
- In the public library environment, anyone can be a user — members of the public, adults, children, students, housewives, literate, neo-literate and even illiterate people, professionals, agriculturists, artisans, planners and policy mpolicymakers on.

Pao mentions that the term users is quite ambiguous. There are several distinct types of users of an information system. Within the context of an organization, there could be:

- Actual users, that is, those who are using the information service at a given time
- Potential users, that is, those who are not yet served by the information service
- Expected users, that is, those who not only have the privilege of using the information service but also have the intention of doing so
- K beneficiary users, that is, those who have derived some benefit from the information service.

- Atherton mentions that three important groups of users of a scientific and technical information system are distinguishable according to the kind of activity in which they are engaged:
- Researchers, in basic and applied sciences
- Practitioners and technicians engaged in developmental and/or operational activities in the various fields of technology and industry, agriculture, medicine, industrial production, communication, and so on.
- R managers, planners, and decision-makers.

These user groups are very broadly defined; the categorization is by no means exhaustive. The list does not include some other user groups, such as students and teachers. There is a lot of cross-classification of users too. For example, a researcher may be at the same time a manager, planner or policy maker.
Guinchat and Menu have employed two objective criteria to define users:
Objective criteria, such as the socio-professional category, specialist field, nature of the activity for which information is sought, and reason for using the information system K social and psychological criteria, such as the users' attitudes and values in regard to information in general and their relation with information units in particular, the reasons behind their particular information-seeking, and their professional and social behavior.
Guinchat and Menou also identified the following broad categories of user based on the two criteria mentioned above:

- Users not yet engaged in active work life, such as students
- Users with a job and whose information needs are related to their work; these users may be classified by the nature of their activity, such as management, research, development, production, or services, by activities in a branch and/or specialist field, such as the civil service, agriculture, or industry, and by level of education and responsibility, such as professional, technical, etc.
- The ordinary citizen requiring general information for social purposes.

4.5.1 TYPES OF INFORMATION NEEDS

Information need is often a vague concept. It is often a result of some unresolved problem(s). It may arise when an individual recognizes that their current state of knowledge is insufficient to cope with the task in hand, or in order to resolve conflicts in a subject area, or to fill a void in some area of knowledge. Before going on to identify the information needs of different categories of users, the following points should be kept in mind:

- Information need is a relative concept; it depends on several factors and does not remain constant.
- Information needs change over a period.

- Information needs vary from person to person, from job to job, subject to subject, organization to organization, and so on.
  People's information needs are largely dependent on the environment; for example, the information needs of those in an academic environment are different from those in an industrial, business, government or administrative environment measuring (quantifying) information need is difficult information need often remains unexpressed or poorly expressed information need often changes upon receipt of some information.

Taylor in the context of library search identifies four major types of information need that lead the user from the state of a purely conceptual need to one that is formally expressed and constrained (by the environment):
Visceral need -+ Conscious need -+ Formalized need —• Compromised need
Were
Visceral need is the unconscious need Conscious need: conscious by undefined need Formalized need: formally expressed need
Compromised need: expressed need influenced by internal and external constraints
Xie suggests that Taylor's work has formed the foundation of several research studies in interactive information seeking and retrieval, including those of Belkin, Kuhlthau, and Ingwersen.
We have already seen that information retrieval system need not be limited to the four walls of any library. There could be information retrieval systems designed to serve a group of users engaged in a specific kind of activity or mission; such information systems are often called information support systems or mission-ooriented information systems. Users of such systems could be students, academics, researchers, planners, policymakers, administrators, and so on, the common thread being that all of them are engaged in a specific area of study or activity or are joined to accomplish a particular mission. They could be part of any organization or institution. For example, in a government information system users may broadly be categorized in accordance with the nature or area of activity, such as education, energy, trade and commerce, and so on. In an industrial environment, users may be corporate, industrialists, or professionals such as engineers, managers, accountants, and so on. The same is true for business and commercial information systems. These information systems may have their own home-grown databases as well as access to one or more CD-ROM and/or online resources.
Thus, we can see that the concept of the user depends on the context in which the information retrieval system is viewed. For instance, in the context of a library environment, we have an idea of the nature and category of users, although their nature, number, nature of activities, and consequently the nature of their information requirements constantly change. The design of information retrieval systems to support users engaged in a specific area of study or activity can be much

more challenging. While much of the information content of the databases contained in a library environment will be bibliographic or reference or textual in nature, in the context of an information support system the information content is factual in nature. Factual data are significantly different from bibliographic data. For example, doctors working in a hospital may need information on patients (related to disease, treatment, tests, medication, and so forth), scientists or policymakers working in a pollution control environment may need data related to the level of pollution by area, by pollutants, by amount and frequency of emission, and so on; the list may go on and on. In their day-to-day activities, scientists, engineers, doctors, administrators, planners, and so forth need information that is factual (not necessarily of a bibliographic or textual type), and when they meet difficulties in carrying out a job, in solving a problem, in taking a decision, and so on, they turn to other kinds of databases containing different kinds of information sources — bibliographic, personal, institutional, and so on.

Some of the most important questions in developing an information retrieval system for supporting users in a specific field of activity, therefore, relate to the identification of actual and potential users of the proposed information retrieval system, the nature of their activities, information requirements, and so on. A user survey can help the information manager to gather information on all these and related points.

An understanding of their users' nature, information needs, information-seeking patterns, and so forth assists an information manager at different levels. At the macro level this knowledge helps such a manager:

- To decide whether to establish an information system, and if so, why, how, and so on.
- To evaluate an existing information retrieval system when:
- Starting a new service.
- Increasing or decreasing emphasis on one or more existing services.
- Optimizing a service.
- Marketing a service, and so on.

  At the micro-level this knowledge will help an information manager to:
- Determine who are the users of an existing or proposed information retrieval system:
- R determine the information needs of each category of users.
- H assess how far the existing system can meet the needs of the user H identify what information sources are to be possessed by the system.
- R determine how the information sources are to be analyzed and recorded.
- Determine the hardware and software requirements, nature and format of the database(s), approach to database design (centralized or distributed), networking requirements, standards, protocols, and so on.
- Determine the communication pattern, user interface and so on

- R determine the output format(s) required, the requirement for repackaging of information and so on.
- R determine the marketing strategies — information products, distribution, pricing and so on.
- R determine the level of staff training, user orientation, training and so on.

4.6 INFORMATION NEEDS IN DIFFERENT AREAS OF ACTIVITY

Several good publications are available that talk about the information needs of various categories of users. For example, the information needs of users in business and industry in general, in product planning and development, and in the establishment as well as in the promotion and management of small-scale industries have been discussed by nee1ameghan, while Atherton discusses the needs of users in the field of scientific and technological research and scott and wootliff discuss the information needs in business environments.

4.7 INFORMATION NEEDS IN SCIENTIFIC AND TECHNOLOGICAL RESEARCH

Atherton identified seven different stages in scientific and technical research and the corresponding information need:
Overall familiarization with the problem and problem statement: This stage requires a general acquaintance with the subject for drawing up a plan and provisional terms for the solutions of the problems of primary and secondary importance. Users need general information on the chosen subject in order to build up an overall idea.

- Gathering scientific knowledge about the subject of study: At this stage the user is engaged in the retrospective searching of the broadest possible scope of the literature without any pronounced critical approach.
- Coordination and interpretation of scientific data: Here the user attempts to make a critical evaluation of the ideas and hypotheses of different authors. The relevance criteria for the information needed are specified at this stage and the volume of information is reduced.
- Formulation of the problem: Statement of the hypothesis and choice of the problem are one of the most important stages in a piece of research. As to the need for information, this is characterized by in-depth analysis rather than broad coverage.
- Proving the working hypothesis: Information requirements at this stage depend on the specifics of the research. The researcher may need a lot of factual data at this stage.
- Statement of conclusions and recommendations: At this stage the user may need to conclude based on their findings and on those available in the

literature. The user may need a good amount of consolidated information at this stage to shed light on precedence and priority aspects.

- Description of the research results: At this stage the user requires information on scientific reporting and documentation. Users may need to check each document consulted for bibliographic and other details for the purpose of documentation.

4.8 INFORMATION NEEDS IN BUSINESS

Scott and wootliff state that there are three major categories of user in a business environment.' 0 the following are those categories of user and indications of the nature of the information they need in their day-to-day activities:

1. Planners and market analysts, who need information on future trends and potential new markets in industries, and on what competitors are doing in their field, in order to plan future strategies for business
2. Service professionals, such as accountants, stockbrokers, bankers, and management consultants, who need information on specific industries, such as:

✓ Who the major players in an industry are?
✓ How they are performing
✓ What their percentage shares of the market are
✓ What their major activities are to advise their clients. They also need to understand a client's business; for example, what external factors or forces — interest rates, foreign exchange rates, political change, etc. — influencean that industry, and
✓ What problems are peculiar to it. And they need to keep track of legislation that is going to affect their own or their client's business, for example company law

3. Corporate finance specialists, who need to identify potential takeover targets in a specific industry and require information on their financial performance.
4. Types of enquiries that are common in a business information unit include requests for:

✓ Information on people, for example, directors of a company R economic indicators and trends of one or more companies
✓ Country risk reports, or information on how to do business in a country, and so forth.

4.9 INFORMATION NEEDS IN ENTERPRISES

Neelameghan identifies the following activities involved in the promotion and management of enterprises for which users may need information:

Formulating objectives of the enterprise Formulating major strategies and policies to meet specific objectives preparing long-range plans Reporting to the stockholders or to the board of management about the results of the enterprise's operations.

Informing employees about the status and performance of the enterprise providing bases and background so that decisions can be made about specific matters as they arise Providing bases for giving pre-action approval Building the background for outside contacts, such as legislators, competitors, and governments Taking decisions about taxes and so forth Keeping abreast of current operations and developments in the business concerned being aware of possible troubles and problems ahead allocating capital resources optimally exercising control over day-to-day operations training staff improving personnel management and public relations.

Information needs of persons working on different aspects of product design, development, and production vary, and this must be borne in mind during the development of an information retrieval system. Neelameghan identifies the information needs of persons concerned with product planning and development, and their respective roles and functions in an enterprise, as follows:

1. planning: long-range forecasting of developments and profits and providing the overall direction for development research: setting up priorities on products and projects on the basis of available funds and time, trouble-shooting, avoiding side-tracking, advising the termination of a project when it is sensed to be unprofitable, and so on F engineering: drawing up specifications, designing and testing prototypes, manufacturing, making adjustments in the final engineering design and models, advising on and implementing continuous improvements in the process of production on the basis of feedback, and so on production: scheduling, formulating process and procedures, testing of equipment, materials requirements, pilot runs, and tests, quality control, and so on.
2. marketing: market analysis, consumer research, forecasting of market developments, market testing, finding solutions to distribution problems, sales promotion, advertising, and so on.
3. public relations: all public relation activities, coordinating responses of the firm to criticisms from outside, building the image of the enterprise, and so on.

4.10 INFORMATION REQUIRED TO SUPPORT COMMUNITY DEVELOPMENT PLANNING

Neelameghan provides a detailed account of the different kinds of information required in the process of community development planning." the following are the main points from neelameghan's account.

The major categories of information that might be required are:

1. Information about the geographical environment r population and demographic information
2. Socio-economic information.
3. Such items of information are collected in several ways: through surveys, census data, maps, community profiles, and so on. Information about the community may consist of:
4. General information on the area or boundary of a village, population, households, literacy rate, birth, and mortality rates, and so on.

Information related to special problems of the community:

1. Agriculture and livestock patterns, inputs, and practices
2. Livestock information
3. Fisheries information
4. Cottage industries information
5. Trade information
6. Information on community structure and facilities, and so on.

Household level information may comprise: Information about members of the household Information on employment patterns, Information on housing conditions and related amenities, Information on assets such as land, livestock, poultry and equipment, Consumer expenditure patterns, Basic health conditions, Literacy rate, education pattern, and so on.

The UN Food and Agriculture Organization (FAO) has recommended some basic items of information that may be needed in community development planning. These relate to information on the following points:

- Agricultural (cultivated or harvested) land
- Agricultural area improved by drainage, irrigation, terracing and so on as a percentage of the total agricultural land
- Production and yield rate of crops
- The intensity of cropping
- The number of livestock species and/or units per economically active person in agriculture.
- Institutional and non-institutional loans per household.
- The percentage of the economically active population in agriculture.

- The percentage of the economically inactive population in agriculture.
- The percentage of areas covered by the size of groups of agricultural holdings or holders.
- Agricultural laborer as percentage of population economically active in agriculture.
- The average wage rate of agricultural laborer.
- The percentage of community heads without land.
- The percentage of households who own their houses (or sites).
- The percentage of households in dwellings which are in good condition.
- The percentage of households with specified facilities, e.g. piped water, sanitation, electricity.
- The primary school enrolment ratio.
- The primary school attendance ratio R the total adult literacy rate.
- The percentage of adult rural population participating in designing, monitoring, and evaluating agricultural and rural development programmes. Neelameghan has discussed the information needs in several specialized activities, for example, community development planning, government and administration, and socio-economic development.

4.10.1 INFORMATION-SEEKING BEHAVIOR OF USERS

Information-seeking behavior or the pattern of using information systems and centers depends on a number of factors. Some of these are closely related to the personal characteristics and traits of users, whereas some depend on the information Centre and the information system concerned. Moreover, the general educational level, awareness of people in a society and overall context are also important determining factors influencing information-seeking behavior. The following are some general pointers that can affect the information-seeking behavior of the individual user:

- The educational and professional background and environment in which the user grew up and/or is currently living.
- Their awareness of, and ability to access, sources of information their relationship with the information unit concerned.
- The information unit's ease of accessibility to the user's working conditions.
- The time available to them for consulting information systems on their hierarchical status and socio-professional position.
- Their personal and professional connections how challenging their job is the amount of competition that exists in their field of activities their past experience.
- How much do they already know?
- How easily they get on with people.

- Their general attitude towards people and organizations.
- How friendly, knowledgeable, and efficient the members of the information unit are the various products and services of the information unit how the user formulates their queries.
- How they make use of the information they obtain how user-friendly the information system is.
- How effective the marketing policy of the information unit is.
- How effective the unit's 'user education', 'user sensitization', 'user orientation', and 'user assistance' programs are?

4.11 ACTIVITES

1. Identify which information search method will best suit your objective, and a plan of operation, and based on terms that appropriately represent the search concept, and features of the retrieval system concerned.
2. Make a complete strategy for the user searching a database that has controlled index languages.
3. Consider you have decided to adopt a model-based approach for information retrieval. Discus the pros and cons of user-centered/cognitive models and system-centered models in the view of your situation.

4.12 SELF ASSESSMENT QUESTIONS

1. Which retrieval model reduces the variants of a word to a single form for retrieval purposes?
2. Which points need to consider identifying the information needs of different categories of users?
3. What kind of skills that the successful pre-search interviewer should possess?
4. What are the important factors that can affect the information-seeking behavior of the individual user?

4.13 REFERENCES

Pao, M. L., Concepts of Information Retrieval, Englewood, CO, Libraries Unlimited,1989.

Atherton, P., Handbaok of Information Systems and Services, Paris, Unesco, 1977.

Guinchat, C. and Menou, M., General Introduction to the Techniques oflnformation and Documentation Work, Paris, Unesco, 1983.

Taylor, R., Question-negotiation and Information Seeking, College & Research Libraries, 29 (3), 1968, 178—94.

Xie, I., Interactive Information Retrieval in Digital Environments, Hershey, IGI Publishing, 2008.

Belkin, N. J., Oddy, R. N. and Brooks, H. M., ASK for Information Retrieval, Part 1, background and theory, Journal of Documentation, 38 (2), 1982, 61—71.

Unit–5

USER INTERFACES AND EVALUATION OF INFORMATION RETRIEVAL SYSTEMS AND EVALUATION EXPERIMENTS

Complied by: Dr. Munazza Jabeen

Reviewed by: 1. Dr. Pervaiz Ahmad
2. Dr. Muhammad Arif
3. Muhammad Jawwad

CONTENTS

INTRODUCTION

The user interface forms an important component of an information retrieval system since it connects the users to the organized information resources. User interfaces perform two major functions: they allow users to search or browse an information collection and they display the results of a search. They also often allow users to perform further tasks, such as sorting, saving and/or printing search results, modifying a search query, and so on. The user interface is therefore the most important component of an information retrieval system that a user can see and interact with. The success of an information retrieval system depends significantly on the design and usefulness of the user interface. Hence a significant amount of research has taken place in the past few decades on the design, use and evaluation of user interfaces to various kinds of information retrieval systems.

OBJECTIVES

After reading this unit you would be able to:

i. Acquire information about the framework for interface design and how the user interface forms an essential component of an information retrieval system?

ii. Know to get the correct visualization technique of the user interfaces by precisely facilitating rapid and uncomplicated communication

iii. Comprehend the essential criteria of user interfaces for browsing and searching.

iv. Understand key evaluation criteria of user-centred design of interfaces

5.1 THE FOUR-PHASE FRAMEWORK FOR INTERFACE DESIGN

Information searching is a complex process. It involves several stages and at each stage a number of actions are taken, and decisions are made. The information retrieval system and the user interface may provide support in performing these actions and in making appropriate decisions. Shneiderman, Byrd and Croft divide the major activities in an information search process into four major phases: formulation, action, review of results and refinement. They propose that this four-phase framework for interface design will provide common structure and terminology for information searching while preserving the distinct features of individual digital library collections and search mechanisms.

ACTION

Usually, a search button needs to be pressed to conduct a search. In some cases, the user just needs to press <CR> to activate the search process. Once the search begins, the user is usually expected to wait until the search process is completed. Sometimes this may take a long time and thus may be quite frustrating. In some cases, the interface prompts the user that the search is being processed; it may also tell the user about the progress of the search. A very appealing method of information searching uses 'dynamic queries' where there is no search button; the result set is continuously displayed and updated as phases of the search are changing.

REVIEW OF RESULTS

Information retrieval interfaces usually offer various choices to the user for viewing results by seeking the size of the display, the display format, and the sequencing of the retrieved items (sorted by author, date, and so on). Some interfaces use different visualization techniques for the display of search results. Some interfaces also use helpful messages to explain the results, for example, commentary on the degree of relevance. Some search results screens show the format of the different retrieved items. Many systems display search results that are sorted in order of relevance, but also provide an additional option(s) for sorting the results by other criteria, for example, alphabetically.

REFINEMENT

Search interfaces provide different facilities for modifying and refining queries. In some cases, users need to reformulate the search statement and conduct a new search, while in other users can refine a search and conduct a new search on the retrieved set. For example, in Dialog search, each search is automatically given a set number, and the user can call any search set and refine the search statement to conduct a search on the previously retrieved set of results. Some information

retrieval systems provide a thesaurus interface to help users formulate or modify queries.

5.2 INFORMATION SEEKING AND USER INTERFACES

User interfaces to information retrieval systems that support information-seeking processes have been widely discussed in the literature. Interface design encompasses what appears on the users' screen, how they view it and how they manipulate it. Functional design specifies the functions that are offered to the user, which include selecting parts of a digital object, searching a list, or sorting retrieved output, obtaining help and manipulating objects that appear on the screen. Most PCs have a user interface that is based on the style derived at Xerox PARC (the Palo Alto Research Center) and made popular on Apple Macs, which use the metaphors of files and folders on a desktop.
Schneiderman, the Guru of human—computer interaction and user interface design, proposes a number of guiding principles for the design of user interfaces.

- Strive for consistency in terminology, layout, instructions, fonts and color.
- Provide shortcuts for skilled users.
- Provide appropriate and informative feedback about the sources and what is being searched for.
- Design for closure so that users know when they have completed searching the entire collection or have viewed every item in a browse list.
- Permit the reversal of actions so that users can undo or modify actions; for example, users should be able to modify their queries or go back to the previous state in a search session.
- Support user control, allowing users to monitor the progress of a search and be able to specify the parameters to control a search.
- Reduce short-term memory load; the system should keep track of some important actions performed by the users and allow them to jump easily to a formerly performed action, for example, to a former query or to a specific result set.
- Simple error-handling facilities to allow users to rectify errors easily; all error messages should be clear and specific.
- Provide plenty of space for entering text in search boxes.
- Provide alternative interfaces for expert and novice users.

Interface design is pivotal to the effective use of an information system, and the application environment of information retrieval systems has its own distinctive needs and characteristics, which need to be understood and addressed in design. Hearst comments that a user interface designer must make decisions about how to

arrange various kinds of information on the screen and how to structure the possible sequences of user-system interactions.
Marchionini provides a description of the essential features of interfaces to support end-user information. seeking and suggests five information seeking functions: problem definition, source selection, problem articulation, result exam- ination and information extraction. He argues that much of the interface work has focused on problem articulation (including query formulation) and that other functions need to be investigated in designing information-seeking interfaces. Marchionini and Komlodi discuss the evolution of interfaces and trace research and development in three areas: information seeking, interface design and computer technology. They provide a brief review of interfaces to online information retrieval systems as well as to the online public access catalogues. They also discuss the new generation of user interfaces influenced by the emergence of the web. They conclude that interface design has become more user-centered, and the trend is toward more mature interfaces that support a range of information-seeking strategies.

5.3 USER INTERFACES AND VISUALIZATION

Since human beings are highly attuned to images, and since visual representation facilitates rapid and easy communication, several visualization techniques have now been applied to the design of user interfaces. Various graphical representation and manipulation techniques are used to represent information on the user screens, although visualization of textually represented information can be challenging.
Users of popular operating systems and common software packages employ a number of visual tools and techniques for day-to-day operations. These include icons, colour highlighting, windows and boxes. The most seen visualization techniques used in interfaces for information access include the following:

- perspective wall: resembles a grey wall folded into three parts and provides a sort of fish-eye view; the center panel provides a detailed view and the two wings provide a contextual view; suitable for information that has a linear structure
- cone tree: provides a fish-eye view by displaying the closer nodes larger and brighter than the farther nodes; suitable for information that has a hierarchical structure
- document lenses: used to focus on one page in a document
- Hyperbolic tree browser: used to show the hierarchical structure of a collection as a hyperbolic tree (for a demonstration from the Universal Library site see www.ulib.org/webRoot/hTree)
- brushing and linking: connects two or more views of the same data such that a change to the representation of one view affects the representation of the other

- panning and zooming: mimics the actions of a movie camera, which can scan sideways across a scene, called panning, and can move in for a close-up or back away to get a more distant view, called zooming
- focus plus context: one portion of the collection is made the focus of attention by making it larger while shrinking the surrounding objects that form the context.

5.4 USER INTERFACES OF SOME INFORMATION RETRIEVAL SYSTEMS

Information retrieval systems vary in terms of design, objectives, characteristics, contents and users. Consequently, many different types of user interfaces can be found. While some of these are quite simple, others are quite sophisticated in terms of design features as well as visualization techniques. In this section we shall look at the user interfaces of some online information retrieval systems.

5.5 INTERFACES FOR BROWSING AND SEARCHING

The two basic modes of access to information in digital libraries are browsing and searching and most information retrieval systems provide facilities for both.

5.6 BROWSE SCREENS

A browse screen is provided to help users browse the entire collection of information in a system. Browsing may be done by the title, subject, or topic (as in the browse screen of Institute Social Sciences that shows browse option by subject, or by a combination of categories such as subject and document type. The basic idea is to allow users to select a heading from a given list. Some information retrieval systems provide taxonomy structure for browsing. A typical example is the Entrez Taxonomy Browser. Entrez is the text- based search and retrieval system used at NCBI for all the major databases, including PubMed, Nucleotide and Protein Sequences, Protein Structures, Complete Genomes, Taxonomy, OMIM, and many others (www.ncbi.nlm.nih.gov/Database/ index.html).

5.7 USER-CENTERED DESIGN OF INTERFACES

A user-centered design of interfaces has been proposed by many researchers, for example, Marchionini and Komlodi; Fox and Urs; Ba1donado; Theng et a/. and Meyyappan, Chowdhury and Foo. Several researchers have proposed information access models to support creativity (see for example, Shneiderman). Shneiderman proposes that user interfaces should be designed such that they support the creativity of users. He proposes the genex framework, which supports creativity through four phases:

1. collect: learn from previous works stored in digital libraries

2. relate: consult with peers and mentors
3. create: explore, compose, and evaluate possible solutions
4. disseminate the results and contribute to the digital libraries.

He further describes eight activities that need powerful interfaces to support creative work. In other words, he proposes eight areas that need the attention of researchers to make the future digital libraries useful for creative work. These are:
Searching and browsing digital libraries: Users should have more control over searching and browsing so that they can make use of their prior knowledge and can retrieve information that supports their creative activities. Since searching is a part of the entire creative process, users should be able to save the search results into the appropriate system or software for future use, for example, on a spreadsheet for further manipulation, as a file that can later be used for consultation with peers, or on a personal notebook for later referral.
Consulting with peers: Users may often consult their peers with new findings or research ideas and information is collected at different stages in the consultation process. Different tools and techniques are also used for consultation. This concerns the design and use of technologies for the interface design, since the appropriate balance of privacy as well as rights to, and ease of, access to information is very important.
Visualizing data and processes: Interfaces that support the visualization of digital library contents are very useful and further works are necessary for a smooth integration of the technologies. For example, the interface should allow the user to view the results of a search using appropriate visualization tools that would help them select the most appropriate results.
Appropriate packages, for example, to a spreadsheet or to a database, and eventually the processed information may be included in a report or presentation.
Reviewing and replaying session histories: Digital library users may like to replay previous sessions to get some new information or to begin from there for a new search session. However, as Shneiderman comments, success in this requires careful user interface and software design to ensure that the results are compact, comprehensible, and useful. Disseminating results: new information may be disseminated to different types of users. The first possible group would be previous and current researchers in the field. Digital libraries should allow users to easily find workers in a field of study. Shneiderman recommends that digital libraries could be Evaluation of information retrieval systems.

5.8 THE PURPOSE OF EVALUATION

Evaluation studies investigate the degree to which the stated goals or expectations have been achieved or the degree to which these can be achieved. Keen gives three major purposes of evaluating an information retrieval system:

1. The need for measures with which to make merit comparisons within a single test situation. In other words, evaluation studies are conducted to compare the merits (or demerits) of two or more systems
2. The need for measures with which to make comparisons between results obtained in different test situations, and
3. The need for assessing the merit of a real-life system.

   Swanson states that evaluation studies have one or more of the following purposes:

   - to assess a set of goals, a program plan, or a design prior to implementation R to determine whether and how well goals or performance expectations are being fulfilled
   - to determine specific reasons for successes and failures
   - to uncover principles underlying a successful program
   - to explore techniques for increasing program effectiveness
   - to establish a foundation of further research on the reasons for the relative success of alternative techniques, and
   - to improve the means employed for attaining objectives or to redefine sub-goals or goals in view of research findings.

5.9 EVALUATION CRITERIA

An evaluation study can be conducted from two different points of view. When it is conducted from a managerial point of view, the evaluation study is called management-oriented; conducted from the users' point of view, it is called a user-oriented evaluation study. Many information scientists advocate that evaluation of an information retrieval system should always be user-oriented — evaluators should pay more attention to those factors that can provide improved service to the users. Cleverdon says that a user-oriented evaluation should try to answer the following questions:

- To what extent does the system meet both the expressed and latent needs of its users' community?
- What are the reasons for the failure of the system to meet the users' needs?
- What is the cost-effectiveness of the searches made by the users themselves as against those made by the intermediaries?
- What basic changes are required to improve the output?
- Can the costs be reduced while maintaining the same level of performance?
- What would be the possible effect if some new services were introduced, or an existing service were withdrawn?

As with any other system, we expect the best possible performance at the least cost from an information retrieval system. We can thus identify two major factors: performance and cost. Now, if we try to determine how we measure the

performance of an information retrieval system we have to go back to the question of its basic objective. We know that the system is intended to retrieve all those documents in a collection that are relevant to a given query while holding back all those documents that are not relevant. The system, therefore, should retrieve — and only retrieve relevant items. The question of relevance thus becomes an important factor. We shall come to this issue shortly. We also want to assess how economically a system performs. The calculation of costs of an information retrieval system is not easy, as it involves several indirect methods of cost calculation.

Lancaster lists the following major factors to be taken into consideration for cost calculation:

- cost incurred per search
- users' efforts involved
- in learning how the system works
- in actual use
- in getting the documents through backup document delivery systems
- in retrieving information from the retrieved documents
- users' time
- from submission of the query to the retrieval of references
- from submission of the query to the retrieval of documents and the actual information.

Several studies have been conducted so far to determine the costs of information retrieval systems or subsystems. Detailed discussions on these studies are available in Roberts and are not discussed here. In this unit, we shall concentrate on the factors relating to the performance of an information retrieval system.

Saracevic mentions that S. C. Bradford was the first person to use the term 'relevance' in the context that it is used today in the field of information science. In the context of the information retrieval system, relevance is a measure of the contact between a source and a destination — between a document and its user. According to Saracevic, many studies in information science have concentrated on determining:

- What factors enter the notion of relevance and
- What relation the notion of relevance specifies.

Both issues require the identification of the performance factors that are the parameters for assessing relevance. In 1997 Mizzaro presented a history of 'relevance' through an exhaustive review of 157 works on the subject: 'it seems clear that the "1959-1976" period is more oriented towards a relevance inherent in document and query: some problems are noted but operationally supposedly negligible. In the "1977—present" period these problems are tackled, and the researchers try to understand, formalize and measure a more subjective, dynamic

and multidimensional relevance.' In 1966 Claverdon identified six criteria for the evaluation of an information retrieval system:

1. Recall: the ability of the system to present all the relevant items
2. Time lag: the average interval between the time that the search request is made and when an answer is provided
3. Effort, intellectual as well as physical, required from the user in obtaining answers to the search requests
4. Form of presentation of the search output, which affects the user's ability to make use of the retrieved items, and
5. Coverage of the collection: the extent to which the system includes relevant matter.

   Vickery identifies six criteria, grouped into two sets:

   Set 1:
   - Coverage: the proportion of the total potentially useful literature that has been analyzed
   - Recall: the proportion of such references that are retrieved in a search
   - Response time: the average time needed to obtain a response from the system.

These three criteria are related to the availability of information, while the following three are related to the selectivity of output:

Set 2:
- Precision: the ability of the system to screen out irrelevant references
- Usability: the value of the references retrieved, in terms of such factors
- As their reliability, comprehensibility, and currency
- Presentation: the form in which search results are presented to the user.

5.10 THE STEPS OF EVALUATION

Lancaster identifies five major steps involved in the evaluation of an information retrieval system:

1. Designing the scope of evaluation
2. Designing the evaluation programme
3. Execution of the evaluation
4. Analysis and interpretation of results
5. Modifying the system in the light of the evaluation results

Step 1 An evaluation study is conducted to determine the level of performance of the given system and to identify those factors that are the reasons for weaknesses.

In other words, an attempt is made to find out the different parameters and their interrelations with a view to assessing their contribution to the overall performance of the system. The first step of an evaluation study entails the preparation of a set of objectives that the given study is going to meet. The purpose and scope of the whole evaluation programme are set at this step. How the evaluation study will be conducted is also considered — in a laboratory-type set-up or in a real-life situation, at what level it will be evaluated — macro evaluation or micro evaluation, and so on. The probable constraints — in terms of cost, staff time, and so on, are also mentioned at this stage. In fact, a detailed plan is chalked out at this stage that forms the basis of the rest of the programme.

Step 2 Once the basic objectives are set and the proposed plans are outlined, the designer goes on to identify the points on which data are to be collected. At this step the parameters on which data are to be collected are determined, and the methodology is proposed. A detailed plan of action is to be prepared which is to be followed for the collection of data. It is also necessary to draw up a plan for the proposed manipulation of data for reaching a conclusion. It may be noted that while conducting an evaluatprogramamme, the designer might need to control some of the parameters of the system. It is therefore necessary that, while preparing the detailed plan of action, the designer points out which parameters are to be held constant during the study and how this is to be done. In most cases, the detailed design of an evaluation programme is prepared by supervisory staff and systems analysts, while the actual evaluation study is executed by other staff members. It is therefore required that the design should be clear at all points. The design should also mark the major caution points where more care is needed to avoid faults.

Step 3 Execution of the evaluation is obviously the most time-consuming step in an evaluation study. The system personnel collect data in a way prescribed at the design stage. In most cases, a repeated number of observations are required to avoid sampling error and bias. Although the evaluator at this stage has to follow the plan of action thoroughly, they may find some interesting features of the system that were not mentioned at the design stage. It is thus important that there should be a communication between the evaluator and the designer at this stage to share any interesting observations that might call for redesign of the evaluation programme.

Step 4 The whole fate of the evaluation programme rests upon the method of interpretation of results and its accuracy. On the one hand the evaluator has a set of objectives of the evaluation programme, and on the other the observations — the data collected on different parameters. Although the methodology for manipulation of the data is determined at the design stage, the evaluator might need to make some changes to arrive at a better conclusion. Once the data have been manipulated in a

suitable way, the evaluator gets a set of results that is to be interpreted in the light of the set of objectives. The evaluator might need to conduct a failure analysis to justify the results and also suggest improvements. Lancaster mentions that the joint use of performance figures and failure analysis should answer most of the questions identified in the objectives of the evaluation.

Step 5 Finally, the retrieval system is modified, if necessary, considering the results of the evaluation study.

5.11 DISCUSSION

While the classical information retrieval parameters, such as recall and precision, have been used in information retrieval experiments for over four decades, applying them — especially recall — in the modern-day online information retrieval evaluation, is a difficult task. Hence researchers have proposed, and experimented on, new retrieval parameters such as relative recall. These are discussed in the following unit.

5.12 ACTIVITIES

1. Identify four- phase framework for your interface design that should provide common structure and terminology for information searching while preserving the distinct features of individual digital library collections.
2. Make a complete strategy for designing a user-interface for your information retrieval system in the view of Shneiderman's guiding principles for the design of user interfaces.
3. Configure the salient features of icons, color highlighting, windows and boxes for the effective visualization of your interface.

5.13 SELF ASSESSMENT QUESTIONS

1. You have developed a user interface for an information retrieval system. Evaluate your interface in the view of following questions:
2. To what extent does the user interface meet both the expressed and latent needs of its users' community?
3. What could be the possible reasons if your interface fails to meet the users' needs?
4. What is the cost-effectiveness of the searches made by the users themselves as against those made by the intermediaries?
5. What basic changes are required to improve the output?
6. Can the costs be reduced while maintaining the same level of performance?
7. What would be the possible effect if some new services were introduced, or an existing service were withdrawn?

5.14 REFERENCES

Lancaster, F. W., The Cost-Effectiveness Analysis of Information Retrieval and Dissemination Systems, Journal of the American Society for Information Science, 22 (1), 1971, 12-27.

Keen, E. M., Evaluation Parameters. In Salton, G. (ed.), The SMART Retrieval System: experiments in automatic document processing. Englewood Cliffs, NJ, Prentice-Hall, 1971, 7W111.

Swanson, R. W., Performing Evaluation Studies in Information Science. In King D. W. (ed.), Key Papers in Design and Evaluation of Retrieval Systems, New York, Knowledge Industry, 1978, 58-74.

Cleverdon, C. W., User Evaluation of Information Retrieval Systems. In King, D. W. (ed.), Key Papers in Design and Evaluation of Retrieval Systems, New York, Knowledge Industry, 1978, 15W165.

Lancaster, F. W., Information Retrieval Systems. characteristics, testing and evaluation, New York, John Wiley, 1979.

Roberts, S. A. (ed.), Costing and Economics of Library and Information Services, London, Aslib, 1984.

Saracevic, T., Relevance: a review of and a framework for the thinking of the notion in information science. In King, D. W. (ed.), Key Papers in Design and Evaluation of Retrieval Systems, New York, Knowledge Industry, 1978, 8W106.

Mizzaro, S., Relevance: the whole history, Journal of the American Society for Information Science, 48 (9), 1997, 81£r-32.

Vickery, B. C., Techniques of Information Retrieval, London, Butterworth, 1970.

Salton, G. and McGill, M. J., Introduction to Modern Information Retrieval, New York, McGraw-Hill, 1983.

Fugmann, R., Subject Analysis and Indexing.- theoretical foundation and practical advice, Frankfurt, Indeks Verlag, 1993.

Unit–6

Online and CD-ROM information retrieval & multimedia information retrieval

Complied by: **Dr. Munazza Jabeen**

Reviewed by: **1. Dr. Pervaiz Ahmad**
2. Dr. Muhammad Arif
3. Muhammad Jawwad

CONTENTS

INTRODUCTION

Online information retrieval involves searching remotely located databases through interactive communication with the help of computers and communication channels. The database can be accessed by the user directly or via a vendor (supplier of online services), in each case through the computer and communication network. The term ‘online retrieval’ can thus be used to indicate the information retrieval services available from producers of databases, or vendors of these databases. Although online information retrieval systems have existed for more than three decades, recent developments in the internet and World Wide Web have brought significant changes and improvements in the online information retrieval environment. This unit discusses the basic concepts of online information retrieval. Computers have traditionally been used to process numeric as well as textual information. However, although text (including numeric figures, tables, and so on) has been the most used medium, information can be communicated by sound, by picture (graphics), and moving images. Human beings have been communicating information in textual form for centuries and libraries and information centers have been engaged in making this kind of information available to the user community.' There are many fields of work that require access to non-textual information. For example, medical professionals need access to X-rays, architects to building plans, ornithologists to bird calls, estate agents to property photographs, and car engineers (and buyers) to photographs and sounds of car engines. In these and in many other fields non-textual is at least equally as important as textual information. With the recent advances in quality and reductions in the price of display and storage technology, computers are being used more regularly for storage and handling of moving images, animation, and sound, in addition to text and numerals.

OBJECTIVES

After reading this unit you would be able to:

1. Understand the online information retrieval systems and their use and how they can be used to indicate the information retrieval services available from producers of databases?
2. Learn to develop online searching services using the basic steps in an online search
3. Understand the differences between CD-ROM VS online databases
4. Learn multimedia information retrieval and how audio can be used for information retrieval

6.1 ONLINE SEARCHING

The phrase 'online searching' was originally used to describe the process of directly interrogating computer systems to resolve requests for information. Now the phrase is used to denote searches that are conducted by means of a local computer that communicates with a remote computer system containing databases. Users can access the database(s) via an online search service provider (also called vendor). The search process is interactive, and the user can conduct the search iteratively until a satisfactory result is obtained.

With the advent of the internet and World Wide Web, the connotation of online searching has changed. Now we can conduct online searches through the worldwide. Web on information sources that are distributed all over the world. For searching these information sources through the web, we can go straight to the web page of the service provider so long as we know the URL (uniform resource locator, or the address of the web page). Alternatively, we can try to locate the information source(s) by searching through the web search engines (the retrieval programs that help us search the web) such as AltaVista and InfoSeek, or through subject directories or gateways (subject directories that can be navigated to reach a particular information source or a group of similar sources) such as Yahoo! and Intuit.

This unit discusses the former type of online service, the traditional online service characterized by a remote online database search service offered commercially by a search service provider or vendor. One major advantage of this kind of online searching is that it is designed to be pay-as-you-go, and therefore each search session can be costed. Another advantage of online searching is its speed and the currency of the data retrieved. Originally online search services were very expensive and could be complex, and therefore intermediaries were needed to help end-users conduct an effective and efficient online search. However, over the year online search services have become less expensive and more user-friendly. As a result, they can now be used by end-users themselves.

6.2 DEVELOPMENT OF ONLINE SEARCHING

The first major online dial-up service was MEDLINE, the online version of MEDLARS, which was followed in 1972 by the offer of commercial online services from Dialog (Lockheed) and ORBIT (SDC). After 1972 many organizations began to offer online databases and search services. By 1975 there were as many as 300 public access databases available from a range of different vendors. 1Initially, the majority of the online databases were used to provide bibliographic references as the output of search session(s) and these were called bibliographic or reference databases. However, for the past few years, more and more databases are becoming available that retrieve actual information rather than

mere bibliographic references. These databases are either full text, where the full texts of documents (including graphics and pictures) are available or databanks that contain machine-readable numerical (often combined with textual and graphical) data.

Rowley identifies three generations of online searching:

- The first generation, from the beginning to 1981, was characterized by dumb terminals, slow transmission speeds, and mostly bibliographic databases
- The second generation, which lasted through the 1980s, was characterized by PCs as workstations, medium transmission speeds, bibliographic as well as full-text databases, and interfaces directed at the end-users
- The third generation, which started at the beginning of the 1990s, is characterized by multimedia PCs, higher transmission speeds, bibliographic as well as full-text databases, and improved user interfaces, help and tutorial facilities.

To this we can add a new, fourth, generation, which started at the end of the 1990s with web access to online search services. Nowadays, users can go directly to the web address of an online service provider whereby they will discover a screen to log in to the service. Web-based online search services, such as Dialog Web, OVID Online, OCLC First Search, provide fast and easy access to online databases with several search and retrieval facilities. The qualities of online search services coupled with the advantages of the world wide web have brought significant developments to online search systems and have made online searching more directed towards end-users.

The growth in the database industry can be interpreted in terms of the number of vendors or service providers, database producers, databases, database records and online searches. There are several publications that regularly record the growth of online information retrieval, the most prominent publication in this field now being the Gale Directory of Databases (2002) and the most prominent author being Martha Williams.

6.3 ONLINE SEARCH SERVICES

There are various components of an online search service:

- information providers or database producers who provide databases to be accessed in an online mode
- a search service provider or vendor, which provides access to the databases and software for conducting the search
- communication links that connect the user with the host and the database(s); nowadays users can communicate with the service providers through the internet
- a local workstation through which the user is linked to the service.

Online search services, or vendors, are those organizations that provide value-added processing to the databases and offer search services. The following are some examples of online search services:

- Dialog (www.dialog.com/about): A pioneer in online search services, Dialog provides online access to over 800 million records in 900 databases in different disciplines.
- OCLC First Search (www.oclc.org/firstsearch): This provides library users with instant online access to more than 72 databases, including these valuable OCLC databases: OCLC World Cat, OCLC First Search Electronic Collections Online.

6.4 BASIC STEPS IN AN ONLINE SEARCH

The steps involved in carrying out an online search vary from system to system. This is because each system has its own custom-built interface, which allows specific types of searches and uses specific operators for different search commands. Nevertheless, the graphical user interfaces used in these systems have made the task of searching reasonably straightforward and the process of searching has been simplified further in the web-based interfaces. These are the basic steps that one needs to follow to conduct an online search.

1. Study the search topic and develop a clear understanding of the information requirement. This is a critical step and depends on a number of factors, such as, the nature and requirements of the user, how well the user can express them
2. Information needs, how much the user already knows, how the user is going to use the information, and so on. This happens before the actual search process begins and is often conducted through a series of dialogues between a searcher and an information intermediary. In the absence of an intermediary, users must clearly delineate their information requirements for themselves.
3. The online service provider as well as their user ID and password.
4. Select the appropriate database(s) to search. This is a critical and often difficult task. The success of a search largely depends on the appropriate selection of the databases.
5. Online search services allow users to select one or more databases to search using the same interface. Most search services allow users to browse through the database categories to select appropriate databases(s). Dialog has a unique facility called Dial Index search (details are given later in this unit), which allows users to see how many times a given search term occurs in a set of chosen databases. This information can guide users to select the appropriate database to conduct the actual search.
6. Formulate search expressions. This is the key part of the job. It may involve

several activities, the first being the selection of appropriate terms and/or phrases. This may require the user to consult dictionaries and thesauri. Once the appropriate search terms and/or phrases are chosen, the search expression must be formulated. At this stage, the user should understand the nature, content, and structure of the chosen database(s) and know which fields are indexed and therefore can be searched. The user also needs to know what search facilities are available, such as Boolean search, truncation, field-specific search, proximity search, and so on, and the appropriate operators. The search operators and syntax for formulating search expressions vary from one search service to the other. Many search service providers have different interfaces for novice and expert users. If the users want to use the expert search interface, which may be command-driven, they have to possess a knowledge of the various search commands and their order of execution.

7. Select the appropriate format for display. Online search services allow users to
8. Select an appropriate format, from several predefined formats, to display the retrieved records. However, there may be charges for the records displayed. For example, when searching Dialog, charges incurred include output and search time costs, as well as internet charges; prices also vary by database.
9. Therefore, one must be very careful in deciding which record(s) to display and in which format to display them. If the option for the display of the full record(s) is chosen, the process may take some time, depending on the network traffic. However, each online search service provides an option for a brief display, which shows the brief details of the output records, and users may select records from this list for a full display.

6.5 FEATURES OF AN ONLINE SEARCH SERVICE: DIALOG WEB

Dialog Web is the web interface to the Dialog online search service, one of the oldest and largest online search service providers, which gives easy access to many databases with:

- company information — directory listings and financial information
- industry information — trends; overviews; market research; specialized industry newsletters and reports; US and international news, including an extensive collection of newspapers and newswires from North America and Asia; and US government news, including public affairs, law and regulatory information
- patents and trademarks — a worldwide collection for research and competitive intelligence tracking
- chemistry, environment, science, and technology — technical literature and reference material to support research needs

- social science and humanities including education, information science, psychology, sociology, and science, from public opinion, news, and leading scholarly and popular publications
- general reference information — people, books, consumer news and travel.

Users can search and retrieve information from all these different types of information sources using:

- Guided Search mode, which does not require knowledge of the Dialog command language
- Command Search mode, which allows experienced users to use the Dialog command language
- database selection tools, which help users pinpoint the right database for a search
- integrated database descriptions, pricing information, and other search assistance
- easy to use forms to create and modify Alerts (current awareness updates).

Dialog search results are available in HTML or text formats. Users have a choice of displaying records or sending search results via e-mail, fax, or postal delivery.

6.6 STEPS IN A DIALOG WEB SEARCH

The first step of a Dialog Web search involves logging in to the system, for which a Dialog account is necessary. The user goes to the Dialog Web site (www.dialogweb.com) and must enter the user ID and password. The log-in screen also provides information about Dialog Web and a preview and search tips. After logging in, the user needs to select the mode of search: Guided Search or Command Search. Guided Search is the default search option.

6.6.1 GUIDED SEARCH

Guided Search is designed for novice to intermediate searchers. The following steps are to be followed for conducting a Guided Search looking for information on digital libraries.

Choose database. To begin the Guided Search, the user clicks the New Search button and chooses from the list of main subject categories. Each category is further divided into focused search topics. For a search on digital libraries, one can select these categories.

6.6.2 CHOOSE A SEARCH OPTION AND CARRY OUT SEARCH

In Guided Search there are two search options:

Targeted Search, which is available in some, but not all, subject categories. It is a

ready-made search form with databases pre-assigned to the form.

1. Dynamic Search, which is available in all the subject categories. The Dynamic Search form is generated based on the category or database that is selected.
2. Dynamic Search has access to many more databases in comparison to the Targeted Search and is more flexible.
3. Targeted Search is the easiest type of search to perform. The user can enter the search word or phrase as 'Words in Title' or as the 'Main Subject'.
4. Dynamic Search is available at various points in the search category selection process or when a user chooses the Quick Functions option in New Search and enters a specific database number. The Dynamic Search capability is available no matter what category or database is picked. In a category with many databases assigned to it, a user can search:

- all the databases together
- a group of similarly designed databases together
- one of the assigned databases individually

If a user has chosen the Dynamic Search option and has decided to conduct the search on all the 12 databases under the 'Library and Information Science' category, the 'Dynamic Search' screen is shown. The Dynamic Search forms also offer the following options:

- Navigation: The search category selections display at the top of the form. To return to a category or option, the user clicks the search category or option name.
- Run Saved Strategy: If a user has already saved a search strategy, it can be run against the selected databases by clicking Run Saved Strategy.

A list of the databases used in the search is displayed at the bottom of the form. The info (i) icon gives more information about the database content and pricing. In the Dynamic Search screen, users can enter a search term or phrase and conduct the search on the subject, author, and descriptor or title field. A search can also be restricted by the year of publication, and the user can browse the list of items by author or year of publication.

6.6.3 DISPLAY SEARCH RESULTS

The search results from a Targeted Search or a Dynamic Search will appear on a Picklist page, which provides a quick view of the records. From the Picklist page users can choose to:

- display specific records in more detailed formats or send records via e-mail or fax, or by post
- rearrange the order in which the records are displayed
- refine the search strategy
- remove duplicate records

- view the prices for all format options
- save the strategy for future use
- create an Alert for automatic updates on the search topic.

After the search has finished processing, the Picklist page will appear. Users can choose to view results by selecting one or more items by checking the boxes and then selecting the display button or can display any one record just by clicking on the hyperlinked title. The format for display is chosen from the 'Format' box and the records are sorted according to a sort criterion chosen from the 'Sort by' box. The search expression can be refined by clicking the 'Back to Search' button, which allows users to edit, add or delete information from the search form.

6.7 COMMAND SEARCH

Command Search is designed for intermediate to experienced Dialog searchers. It provides complete command-based access to Dialog 's extensive collection of databases. Users are expected to be familiar with the various Dialog commands - when using Command Search. Additional features include built-in tools such as Bluesheets (database descriptions) and pricing information, database selection assistance to help pinpoint the right databases for a search and easy to use forms to create and modify Alerts (current awareness updates). The Command Search main page allows users to begin inputting Dialog commands immediately. A Command Search contains:

a textbox for entering Dialog search commands
a Submit button that sends the command

- a Previous button that displays your most recent command entries.

The main page has links to the Databases feature, product support information, and Guided Search. Users can move between Guided Search and Command Search. Steps for conducting a Command search can be summarized as follows.

STEP 1 Choose databases: Dialog Web simplifies database selection by arranging the databases by subject in the Databases feature. Users can select one or more databases by checking in the Database box. However, if users are not sure which database(s) to select, they can choose the Dialog Index option. This is particularly useful when users do not know which databases to search, or when they want to carry out a comprehensive search and cover everything on a topic. Dialog Index is a master index to most of the Dialog databases, and it allows users to compare the number of records retrieved from a group of databases. After selecting a database, users must search the databases to view the records. They can click 'Begin Databases' to enter the files that they have checked and to run the same strategy or may choose the

database(s) to search by entering the file numbers and even changing their search strategy in the command line.

STEP 2 Choose a search option to carry out search: Once the databases are chosen, the Dialog Command Search page appears. It can also appear:

- after log-in if it is set as the default
- when the Command Search link from the main Guided Search page is clicked R when the Begin Databases button from Databases is clicked for browsing. The appropriate BEGIN or 'b' command is inserted in the command line automatically when a search has been made in Databases and one or more databases have been selected. Users can add the CURRENT command to their BEGIN statement by typing in 'current' after the command. This allows them to search the current year and one year earlier and narrows the search results at the beginning. Then they click the Submit button or press the ENTER key on the keyboard to start the search.

Any Dialog command followed by the search term(s) should be entered in the search box. The terms when looking for information on digital libraries might be S digital (w) libraries.

STEP 3 Add to a search: The search can be refined by including 'electronic libraries through the following expression:

1. S (digital or electronic) (w) libraries: This search statement will retrieve records on electronic as well as digital libraries. The search statement retrieves those records where digital and libraries, or electronic and libraries, occur next to each other in the same sequence. More records are retrieved by truncating the search term libraries as follows:
2. S (digital or electronic) (w) library: Various other modifications may be made by using appropriate search commands; for example, limiting the search to one or more fields or limiting the results to a language, year of publication, and so on.

STEP 4 Displaying records: A search history of all of the sets appears and users can view some of the records. It is a good idea to display a few records in 'free' format before displaying the records in the long or full format. To display records users can choose a format from the drop-down list and click Display for the appropriate set. Formats determine the amount of information to be displayed for each record. The Format list box lists the basic format options: free, short, medium, long, full and KWIC. It is possible to indicate the number of records to display; the default is 10 and a maximum of 99 records can be specified. There is an option of using a Type command (see below for details) to display records or From Each together with the Type

command, in order to search more than one database.

6.8 DIALOG SEARCH OPERATORS

Dialog offers several search features, such as Boolean search, proximity search, truncation, field-specific search, limiting search, and so on. The various search features of Dialog and the corresponding operators.

6.9 CD-ROM DATABASES

Since the early 1960s, the computer industry, being aware of some of the limitations of magnetic storage media, has devoted a great deal of research and development effort to the investigation of alternative high-density recording techniques, particularly optical data storage systems. With magnetic media the presence of dust can cause loss of information, but with holographic storage systems dust or scratching results only in a slight loss of resolution as the data are recorded throughout the entire recording medium. Also, for retrieving data on magnetic media, extreme accuracy is required in the location of the read head, but the holographic system is very tolerant of positional inaccuracies.
CD-ROM is one of a family of compact disc formats, the best known of which is compact disc digital audio (CD-DA, commonly known as the CD), which was jointly developed by the Sony and Philips companies and announced in 1980. The discs have the same physical dimensions and composition. In 1983 Sony and Philips announced a standard for the CD-ROM format, a logical extension of the CD.

Table 6.A Some essential Dialog commands

gamman0	**Example**	**Description**
Begin or B	B 202 B 1, 202	Opens one or more databases for searching
Current	S 1 current2	Used to narrow the search to records from the most current year(s) within a file. You can specify the number of years
Find or F	F internet F world wide web	Used in place of the select or s command to conduct a search. The difference with select is that it does not need a proximity operator to search for a phrase
Help or?	Help field 1	Produces help. For example, the command shown here will show the list of fields for Dialog file
Save	Save <name>	Saves the entire search strategy since the last begin command in a file with the given file name
Select or s	S digital libraries	The most essential Dialog command: it is necessary to conduct any search

Select Steps or ss	SS internet and information	Creates a set for each search term/phrase and one for the entire search. This is useful when a multi-word search term is given; later on, you can just call the set with any constituent term to conduct another search
Oft	Sort s1/all/au,ti	Sorts the results of a search set (set1 for the given example) by one or more sort keys (here author and title). Each database has a list of sort keys that can be used. You can click on the 'Sort' button to get a list of sort keys for a given database
Rank	Rank de,id	Conducts a statistical analysis on the existing search set. Dialog extracts the specified fields from the record and lists them in a ranked order

Technology, which made it capable of storing textual data. CD-ROMs and audio CDs are both mass-produced using the same physical mastering and replication processes the main difference between them is that additional error detection and correction features are required for accurate retrieval and representation of data on a computer screen.

Optical storage devices were developed in several parallel tracks geared for different sectors of the market. The range of optical media may be divided into three major functional groups:

- read-only optical media
- write-once optical media
- erasable/rewritable optical media.

6.10 CD-ROM TECHNOLOGY

Compact disc technology owes much of its success to the development of technical standards that have been accepted worldwide. Compact disc standards began wit the CD-audio product; the document that specifies the physical and recording characteristics is known as the 'Red Book'. This unit discusses some of the basic technological issues related to CD-ROM technology.

6.11 ACCEPTED STANDARDS

For the data stored on a compact disc to be read by any computer operating system and any CD-ROM drive, the data must be stored in a standard form. In 1985 several major companies involved in producing CD-ROMs agreed to develop a standard for a logical file structure that would be acceptable to a wide range of operating systems. This became known as the High Sierra Standard (after the name of the hotel High Sierra in Las Vegas, Nevada, where the company representatives met), and was later accepted as ISO 9660.

A CD-ROM file management system, therefore, is designed to allow users to view the disc as a collection of files. A complete CD-ROM file management system comprises three major components

- the structure or logical format of data

- the software that writes the data in that format (the origination software)
- the software that reads and translates the logical format for use (the destination software).

The logical format of the CD-ROM is concerned with determining where to put the identifying data on the disc, where to find the subdirectories or directories of files on the disc, how the directory is structured, whether subdirectories are supported, how many files can be stored on a CD-ROM, the performance cost of storing a large number of files, how large an individual file can be, whether files can span multiple volumes and whether files must consist of sequential consecutive sectors. The logical format is broken into two distinct structures: the volume table of contents (VTOC) and the directory structure. The VTOC contains information about the disc, including the location of the disc directory. When the file- manager begins reading a disc, it reads the VTOC before anything else. The directory structure specifies the exact locations of the files on the disc.

A number of groups have been involved in the formation of a CD-ROM logical standard, including:

- the High Sierra Group
- the Information Industry Association
- the American Library Association and the Library and Technology Association
- the Optical Disk Forum.
- the American National Standards Institution (ANSI). R the National Information Standards Organization (NISO)
- the International Standards Organization (ISO)
- the European Computer Manufacturer's Association (ECMA).

6.12 CD-ROM VS ONLINE DATABASES

CD-ROM technology has been in existence since the mid-1980s, while online information retrieval systems have been around for 20 years longer. When CD-ROM technology started becoming popular, many people considered it as an alternative to online information retrieval systems. As an information management tool, CD-ROM offers a potentially attractive and cost-effective alternative to time-sharing systems which provide remote access to databases. Major differences between the CD-ROM and online database options.

6.13 COMMON SEARCH FEATURES AVAILABLE IN CD-ROM DATABASES

Basic information on CD-ROM discs is available in a number of published documents. The following search features are commonly available in CD-ROM databases although, as we shall see later in this unit, the search syntax, operators, and so on, may vary from one CD-ROM to another.

6.14 MULTIMEDIA INFORMATION RETRIEVAL

Multimedia systems use information and communication technologies for the integrated storage and retrieval of information in the form of numbers, text, images, audio and video. Multimedia information has some specific characteristics that makes it distinct from textual information; thus, multimedia information retrieval systems differ from conventional text retrieval systems. Early works on image retrieval, which were based on the textual annotation of multimedia documents, began in the late 1970s, more advanced multimedia information retrieval research.

6.14.1 AUDIO INFORMATION RETRIEVAL

Categories of audio recordings in a collection can vary from natural sounds such as animal cries to human speech, music, and so on. Some parts of the sound recording may be desirable while others may not, for example extraneous noise in the background. Speech and music retrieval have become prominent areas of research over the past few years.

6.14.2 SPEECH RETRIEVAL

Speech is one of the most used mediums of human communication. The conventional tools for capturing and playing speech and audio information has a common problem: the recorded speech or audio material has to be listened to sequentially. Audio equipment provides mechanisms for going backward and forward in a recorded message, but it is difficult to retrieve a particular segment of a speech from a long-recorded voice. Conventional text retrieval techniques may be applied to voice retrieval easily if we could transcribe spoken audio documents. A perfect automatic speech recognition (ASR) system that can efficiently transcribe spoken audio documents would be an ideal solution for speech retrieval. Hidden Markov models (HMM) form the backbone of ASR systems. An HMM is a statistical representation of a speech event such as a word.3 Model parameters are trained on a large corpus of labelled speech data. Once a trained set of HMMs is generated, query speech can be matched to find the most likely model sequence (the recognized words).

6.14.3 MUSIC RETRIEVAL

Although the first published work on music information retrieval (MIR) appeared long ago very little research has so far been done on it. Byrd and Crawford' have reviewed research on MIR and observe that it is still a very immature field. Interest in this area has been slow to develop, as is evident from the literature available (Bainbridge et a/.Downie and Nelson; Lemstrom, Laine and Perttu;Tseng;Uitdenbogerd and Zobel, Wiseman, Rusbridge and Griffin).

Music information consists of seven facets:

- pitch: a quality of sound that is related to the frequency
- tempo: information concerning the duration of a musical event
- harmony: related to the attribute of music; a harmony occurs when two or more pitches sound at the same time
- timbre: an attribute related to the tone, which brings about the aural distinction between a note played by two different instruments
- editing: related to the performance instructions such as fingering, ornamentation, and articulation
- text: related to the lyrics, symphonies, and so on
- bibliography: information about the composer, performer, title of the piece, publisher, and so on.

Downie identifies two major types of MIR systems:

- Analytic or production systems, which are intended for musicologists, music theorists, music composers and music engravers; these systems focus on a number of facets of music
- Locating MIR systems, which are concerned with access to musical works; in addition to the bibliographic keys, these retrieval systems use timbre and harmonic features of music.

Query-based music retrieval relies on similarity matching between the query and the stored music. Archives of MIDI (Musical Instrument Digital Interface) files, which are score-like representations of music, are used for music retrieval. Most MIR systems, such as those provided by the search engines, use text-based retrieval techniques. For example, AltaVista music search allows users to search by the name of the artist, title of the song, and also by file types, such as MP3 (Moving Picture Experts Group Layer 3 Audio), WAV (Windows Wave), Windows Media, Real or other file types. Some digital libraries provide access to digital music. One prominent example is the New Zealand Digital Library (NZDL). Users can search for music in the NZDL that allows music retrieval by particular notes and keyword and title. Users can search for particular notes and/or words that appear in the music document from the search page. Music may be monophonic, when only one note sounds at a time, or polyphonic.

6.14.4 IMAGE RETRIEVAL

The use of images in human communication began long before modern civilization; paintings carved on the walls of caves and ancient architecture testify to that. The use of computers for processing images can be traced back to 1965 with Ivan Sutherland's Sketchpad project, which was not viable given the limited computing resources and high costs associated with them at that point in time.21 However, image processing and retrieval activities began in the 1980s, and became an active area of research interest since the creation of the web in the early 1990s.

6.14.5 IMAGE RETRIEVAL QUERIES

Image retrieval can be based on metadata (such as the creator, date, or location), associated text including the human-assigned descriptors, or image characteristics such as colour, texture and shape. User queries about images may vary depending on the nature and need of the user as well as the nature and content of the image collections they are searching. Some of these queries may be based on one or more attributes of the images, for example, 'show me the images of Sept 11' or 'find images of F-16 fighter planes built 1990 onwards', while other queries may describe the content of the images in some detail, for example:

- display illustrations that may or may not be described properly in words, for example, 'show me all the images of butterflies with a particular [described] texture of colour on the wings' or 'show me a picture of sunset on a golden beach [of Malaysia, say] where the sky is appears to take a particular colour [golden, say]'
- display all the images of a particular characteristic, for example, 'show all the radiology images of patients with a particular [named] disease'

6.15 DISCUSSION

Multimedia information retrieval has a tremendous potential in different areas. However multimedia information retrieval, especially content-based retrieval, is a very complex area, and compared with the history of text retrieval, multimedia information retrieval is relatively new. Most current research in this area is cconcerned with many multimedia retrieval systems are now available some of which were born as an outcome of research projects, while others came out of commercial interests. Users can now search for large collections of audio, images and video through the web and digital libraries. More complex and sophisticated applications of multimedia, especially image and video retrieval, can be seen in the security and surveillance applications.

6.16 ACTIVITIES

1. Identify the traditional online service characterized by a remote online database search service offered commercially by a search service provider or vendor.
2. Your organization has decided to move from CD-ROM databases to ON-LINE databases. Prepare complete strategy to accomplish the job.

6.17 SELF ASSESSMENT QUESTIONS

1. Consider your library has decided to enable an online search service to provide value-added processing to the databases. What kind of basic steps and features you consider in offering such a service?
2. Which multimedia type would you prefer for information retrieval, and why?
3. Suppose your system has intermediate to experienced dialog searchers, which search method you prefer, and why?

6.18 REFERENCES

Walker, G. and Janes, J., Online Retrieval. a dialogue of theory and practice, Libraries Unlimited, 1993.

Rowley, J., The Electronic Library, 4th edn, London, Library Association Publishing, 1999.

Gale Directory ofDatabases, Vol. 1. Online databases 2003, Vol 2: CD-ROM, disketle, magnetic tape, handheld and batch access database products 2003, Gale, 2002.

Forrester, W. H. and Rowlands, J. L., The Online Searcher's Companion, London, Library Association Publishing, 1999.

Large, A., Tedd, L. A. and Hartley, R. J., Information Seeking in the Online Age. principles and practice, London, Bowker-Saur, 1999.

Chowdhury, G. G. and Chowdhury, S., Searching CD-ROM and Online Information Sources, London, Library Association Publishing, 2001.

Hendley, T., An Introduction to the Range of Optical Storage Media. In Oppenheim, C. (ed.), CD-ROM: fundamentals to applications, London, Butterworths, 1988, 1-38.

Hanson, T. and Day, 1. (eds), CD-ROM in Libraries.' management issues, London, Bowker, 1994.

Dunlop, M. D. and van Rijsbergen, C. J., Hypermedia and Free Text Retrieval,Information Processing and Management, 29 (3), 1993, 287-98.

Long, F. L, Zhang, H. and Feng, D. D., Fundamentals of Content-based Image Retrieval. In Feng, D. D., Wan-Chi, S. and Zhang, H. (eds), Multimedia Information Retrieval and Management. technological fundamentals and applications, Springer, 2003, 1-26.

Bertino, E., Catania, B. and Ferrari, C., Multimedia IR: models and languages. In Baeza-Yates, R. and Ribeiro-Neto, B. (eds), Modern Information Retrieval, New York, ACM, 1999, 325—43.

Foote, J., An Overview of Audio Information Retrieval, Multimedia Systems, 7, 1999,2-10.

Olivetti: Video Mail Retrieval Using Voice, http://mi.eng.cam.ac.uk/research/Projects/vmr.

Kassler, M., Toward Musical Information Retrieval, Perspectives of New Music, 4 (2), 1966, 59-67.

Kassler, M., MIR — a simple programming language for musical information retrieval. In Lincoln, H. B. (ed.), fire Computer and Music, Ithaca, NY, Cornell University Press, 1970, 299-327.

Unit–7

Hypertext And Markup Language and Web Information Retrieval

Complied by: **Dr. Munazza Jabeen**

Reviewed by: **1. Dr. Pervaiz Ahmad**
2. Dr. Muhammad Arif
3. Muhammad Jawwad

CONTENTS

Page #

INTRODUCTION

In either situation, the user may be referred to one or more places for further information about the term (its origin, application, and so on). It is very possible that the inquisitive user may soon lose track of the path that has been traversed and will be lost in the 'jungle of information'. These problems typically occur due to the linear structure of documents, which does not allow users to navigate freely through different parts of the same or different documents. A non-linear documents/text structure allows the user to jump from one place in the text to another: this non-linear arrangement of textual material is called hypertext, where the term hyper means 'extension into other dimensions' converting text into a 'multidimensional space'.

The introduction and growth of the World Wide Web (WWW or simply the web) have brought significant changes in the way we access information. Simply speaking, the web is a massive collection of web pages stored on the millions of computers across the world that are linked by the internet.' The development of the web began in 1989 by Tim Berners-Lee and his colleagues at CERN (European Laboratory for Particle Physics in Geneva). They created a protocol, called the Hyper Text Transfer Protocol (HTTP), which standardized communication between servers and clients. Their text- based web browser was made available for general release in January 1992. The web gained rapid acceptance with the creation of a web browser called Mosaic, which was developed in the USA at the National Center for Supercomputing Applications at the University of Illinois and was released in September 1993

OBJECTIVES

After reading this unit you would be able to:

1. Understand the hypertext reference models and hypermedia systems
2. Determine open hypertext and hypermedia system services for the world wide web.
3. Understand various markup languages for moving from traditional to web information retrieval.
4. Comprehend to evaluate Web information retrieval

7.1 HYPERTEXT

7.1.1 THE HISTORY OF HYPERTEXT

Since the mid-1980s, there has been an explosion in interest in hypertext, along with the development of many hypertext systems. Indeed, within a span of ten years or so hypertext (and hypermedia) has brought tremendous changes in the handling and dissemination of information. However, the concept of hypertext has not been with us for much longer than 15 years.

The origin of the basic concept of hypertext and hypermedia goes back more than 50 years. In 1945 Vannevar Bush proposed a non-linear structuring of text that would correspond to the associative nature of the human mind. Although he did not use the term 'hypertext', he described a machine, which he referred to as 'Memex' that could be used to browse and make notes in a voluminous online text and graphics system. Memex would contain a large library of documents, photographs, and sketches. The idea was that Memex would have several screens and a facility for establishing a labelled link between any two points (or nodes) in the library.

7.1.2 HYPERTEXT: DEFINITION AND MEANING

In its literal sense, the term hypertext implies extra dimensions to text. In practice, the term is often used to describe a computer program that allows a person to browse a document by deliberately jumping from text block to text block. According to Parsaye ef al., hypertext is a tool for building and using associative structures. While a normal document is linear and is usually read from beginning to end, the reading of hypertext is open-ended: one can jump from one point to another as desired. The nearest thing to a hypertext that most people are familiar with is a thesaurus. It too is not normally read from beginning to end. Each time a thesaurus is consulted, it is entered at a different location based on the word used to initiate the search, and once the sought term is located, there are pointers that lead the user to other parts of the thesaurus to get more information on the terms related to the sought term. Hypertext can be thought of as an enriched thesaurus where, instead of links between words, links between documents and text fragments are available.

7.1.3 COMPONENTS OF HYPERTEXT

Hypertext retrieval systems are the products of an emerging technology that specifies an alternative approach to the retrieval of information from machine-readable full-text documents.

Hypertext systems provide the facility for any relationship existing between two document representations, or nodes, to be represented by a link. Searches can retrieve individual nodes successively by activating the links between them. Such links may be created by manual or automatic means.

- A document retrieval system may be identified as a hypertext system if its components include the following:
- structural component, consisting of a database of document representations in which the relationships between documents are explicitly represented, such that the document representations and relationships between them together form a network structure
- functional component consisting of a retrieval mechanism of a type that is:
- navigational — it allows users to make decisions at each stage of the retrieval process as to the object(s) that should be retrieved next
- browsing-based — it allows users to search for information without their having to specify a definite target.

7.1.4 HYPERTEXT REFERENCE MODEL

In 1988, with a view to avoiding the inconsistencies of the different approaches to the development of hypertext systems, a workshop was organized at the Dexter Inn in New Hampshire, to achieve a consensus on basic hypertext systems. The workshop came up with a model, called the Dexter hypertext reference model.
The focus of the model has been laid on the storage layer, which consists of a set of components. The interface between the storage and the runtime layers includes the presentation specification, which determines how the components are presented at runtime. For example, these specifications might include information on screen location and size of the window. The within-component layer corresponds to individual applications; its interface to the storage layer is via anchors which consist of an identifier, which can be referred to by /in/cs, and a value that picks out the anchored part of the material. The runtime layer is responsible for handling links, anchors, and components at runtime.

7.1.5 HYPERMEDIA SYSTEMS

Both hypertext and hypermedia systems provide a way of representing and managing information in a flexible non-linear way that is appropriate for many multimedia applications.4 Hypermedia can simply be defined as the creation and representation of links between discrete pieces of different kinds of data — text, numbers, graphics and/or sound. In other words, in a hypermedia system a node may contain text, graphics, animation, images or sound that can be controlled, presented, and edited on a computer.
Arnets and Bogaerts define hypermedia as:
- the hyper-representation of textual and non-textual information
- style of building systems for information representation and management around a network of multimedia nodes connected by typed links
- generic approach to constructing non-linear computer-supported materials

- the flexible linking together of similar or different types of information.

The hallmark of any hypermedia system is its capability to link together related forms of information in a flexible and easily adaptable manner.
There are four ways of working with hypermedia systems:

- hypermedia as a system
- hypermedia as an interface

7.1.6 OPEN HYPERTEXT AND HYPERMEDIA SYSTEMS

This is a major issue for hypertext research. We have already noted that hypertext provides links among concepts and documents. This facility can be used, in conjunction with groupware technology, to create texts with input from different authors at different places or at different points in time. Groupware is software that helps people directly communicate with one another. Groupware and hypertext complement one another in developing what is known as group text, or open hypertext, which allows a group of people to create and access linked text with a provision for communicating directly with one another. The first group text computer system was the augmentation system of the 1960s.From the 1990s, the concept of the group text or open hypertext system is of central importance to researchers: these systems combine simultaneous video, audio and text, and provide decision support.

7.2 MARKUP LANGUAGES

Markup languages help us mark the specific sections of the items with standard codes, which can be interpreted by computer programs to take specific measures, for example to take measures for appropriate appearance of the encoded text (in bold or colour, say), or to extract a specific portion of the item, say the title, keywords or abstract, to store it in a database for further processing. Several markup languages have been developed to serve different purposes. In this unit we will briefly discuss SGML (Standard Generalized Markup Language), HTML (Hypertext Markup Language, the language of the World Wide Web), XML (extensible Markup Language) and XHTML (extensible Hypertext Markup Language).

7.2.1 SGML

SGML was accepted as a standard in 1986 (ISO 8879:198614). This standard was created to provide a set of rules that describe the structure of an electronic document so that it may be interchanged across various computer platforms. SGML also allows users to:

- link files together to form composite documents
- identify where illustrations are to be incorporated into text files

- create different versions of a document in a single file
- add editorial comments to a file
- provide information to supporting programs.

7.2.2 XML

While SGML is too complex and resource-intensive to encode and cannot be processed as it is by the web browsers, and HTML is too simple and only tells the browser how to present an element or how to link to another item, XML aims to offer the best of both worlds. XML is a simple and flexible text format derived from SGML (ISO 8879). Originally designed to meet the challenges of large-scale electronic publishing, XML is also playing an increasingly important role in the exchange of a wide variety of data on the web and elsewhere. It contains a set of rules for designing text formats that let users structure their data.
Development of XML started in 1996 and has been a W3C Recommendation since February 1998. The designers of XML simply took the best parts of SGML, guided by the experience with HTML, and produced something that is powerful and vastly more regular and simpler to use.

7.2.3 XHTML

During 1999 HTML 4 was recast in XML and the resulting XHTML 1.0 became a W3C Recommendation in January 2000. XHTML is the successor of HTML, and a series of specifications has been developed for XHTML. The XHTML family document types are all XML-based and ultimately are designed to work in conjunction with XML-based user agents.

- XHTML 1.0 is specified in three ‘flavors’ (www.w3.org/MarkUp):
- XHTML 1.0 Strict — to be used to get a clean structural markup, free of any markup associated with the layout; this can be used together with W3C’s CSS to get the font, color and layout effects desired

7.3 WEB INFORMATION RETRIEVAL

7.3.1 TRADITIONAL VS WEB INFORMATION RETRIEVAL

Web information retrieval is significantly different from traditional text retrieval systems. These differences mainly stem from a number of typical characteristics of the web such as its distributed architecture, the variety of information available, its growth, the distribution of information and users, and so on. Several researchers have discussed the uniqueness of web information retrieval (see for example, Bharat)
This section discusses some of these issues with a view to highlighting the complexities of web information retrieval.

1. *Distributed nature of the web*: Web resources are distributed all over the world, so complex measures are required to locate, index, and retrieve them. The fact that the computers that are interconnected have different architecture, and the information resources are created using different platforms, software, and standards makes the matter more complex. Most text retrieval systems deal with a set of information resources that is several times smaller in volume than the web. In addition, text retrieval systems usually deal with a set of documents that have been created using a set of standards — hardware, software, and processing standards. When OPACs retrieve distributed information, they use several standards to process it, such as the MARC formats, and to index it, such as Z39.50. No such uniform standard is used for the creation and processing of web information resources.
2. *Size and growth of the web*: The growth of the web has become more and more rapid. The processes of identifying, indexing, and retrieving information become more complex as the size of the web, and hence the volume of information on the web, increases. Conventional text retrieval systems have to be tested and modified to make them suitable for handling the large volume of data on the web.
3. *Deep vs. the surface web*: Information resources on the web can be accessed at two different levels. While millions of web information resources can be accessed by anyone a lot of information is accessible either through authorized access (information that is password-protected, say) or can be generated only by activating an appropriate program. Researchers call the former 'the surface web' and the latter 'the deep web', with a note that the deep web is several times larger than the surface web
4. *Type and format of the documents*: Text retrieval systems deal with textual information only; the web contains a much wider variety, from simple text to multimedia information, and a variety of data and documents. Again, these information resources appear in a variety of formats thereby making the task of indexing and retrieval more complex.
5. *Quality of information*: Since anyone can publish almost anything on the web, it is very difficult to assess the quality of information resources. As opposed to conventional text retrieval systems, which deal with published information resources that have some quality control, web information retrieval systems must deal with many uncontrolled information resources.
6. *Frequency of changes*: Web pages change quite frequently. This is in sharp contrast with the input of conventional text retrieval systems, which deal with relatively static information. Once an information resource is added to a text retrieval system it does not change its content; at the most the entire document is removed from the system. Keeping track of the changes in the

millions of web pages and making necessary changes in the information retrieval system

7. *Is a major challenge*. Another major problem with the web is that the resources (web pages) often move. This information needs to be tracked by the retrieval system to facilitate proper retrieval.
8. *Ownership*: Information resources that are accessible through the web have different access requirements: some information can be accessed and used freely; others require specific permission or access rights, often through payment of fees. Identifying the rights to access is a major challenge for web information retrieval.
9. *Distributed users*: Most text retrieval systems are designed to meet the information needs of a specific user community. Hence text retrieval systems usually have an idea of the nature, characteristics, information needs, search behavior, and so on of the target user community. Web information is in sharp contrast with this. Ideally the users of an information resource on the web may be anyone, located anywhere in the world. This imposes a significant challenge since the designer of a web information retrieval system will have no idea about the target users, their nature, characteristics, location, information search behavior, and so on.
10. *Multiple languages*: Since the web is distributed all over the world, the language of information resources as well as users varies significantly. An ideal web information retrieval system should be able to retrieve the required information irrespective of the language of the query or the source information. This diversity of language poses a tremendous challenge for web information retrieval.
11. *Resource requirements*: A massive number of resources are required to build and run an effective and efficient web information retrieval system. The matter is worsened by the fact that there is no single body that would fund for these resources, and yet everyone wants a good information retrieval system for access to web information resources.

7.3.2 WEB INFORMATION: VOLUME AND GROWTH

Since its inception, the internet has grown at an unprecedented rate. Latest figures show that about a quarter of the world's population now has internet access, but the growth has not been uniform, especially in the developing countries. However, countries such as China, Russia, Brazil and India are catching up very fast.

7.3.3 WEB INFORMATION RETRIEVAL: ISSUES AND CHALLENGES

As stated earlier in this unit, the web is characterized by many distinct features that have implications for the retrieval of information. The web is a platform where

anyone from anywhere can publish virtually any information, in any language or format. In other words, information published on the web may be peer-reviewed, as

7.3.4 ACCESS TO INFORMATION ON THE WEB: THE TOOLS

A user can get access to any website by entering the uniform resource locator (the address of a website) on the browser. A web browser, such as Microsoft Internet Explorer, is a computer program, an essential tool for getting access to the web. A web browser performs two major tasks:

- It knows how to go to a web server on the internet and request a page so that the browser can pull the page through the network and into your machine.
- It knows how to interpret the set of HTML tags within the page to display the page on your screen as the page's creator intended it to be viewed.

7.3.5 WEB INFORMATION RETRIEVAL: EVALUATION STUDIES

Information retrieval through search engines is characterized by simple and intuitive search interfaces that produce large answer sets: high recall and low precision. Often users find it difficult to understand what criteria were used to produce and rank the results because even the high-ranking results do not seem to be relevant to the query. The traditional online information retrieval services were designed to work primarily on structured text databases, and they were designed with a target user community in mind. Information retrieval services on the web deal with text as well as multimedia information resources that are linked with other documents, and there is no target user community as such. A different context for information retrieval emerges from this environment, representing a more 'popular' use of information retrieval, characterized by a broader audience, different document collections and different search models. Researchers have noted that the retrieval characteristics of web information services are designed for general audiences, and they are different

7.3.6 INFORMATION SEEKING ON THE WEB

Information seeking on the web is a topic of increasing interest in many disciplines. Seeking information and tracing relevant information on the web is a complex task because the same information is diffused, appears in various forms and is available through different channels. Karr notes that information search strategies seemed to form a spectrum of developmental sophistication.

7.4 DISCUSSION

Over the past few years, the web has grown rapidly, and has influenced all sections of society; most importantly it has brought a paradigm shift in the ways we publish, organize, seek, and retrieve information. Consequently, web information retrieval has become a major area of research and business, and there is rapid growth and huge competition among the various web search tools. Google is the biggest player in the search engine market and holds almost three-quarters of the marker share in web searching, although there are many other big and small players in the market. The web has brought several new challenges in information retrieval, and companies are investing huge amounts of resources in developing new tools, technologies, and standards for building improved and more sophisticated web search tools. As well as the computational and algorithmic approaches developed and adopted by web search engines, a new group of web search tools has appeared over the past few years, known as social search engines or social search tools. These take many forms, 'ranging from simple shared bookmarks or tagging of content with descriptive labels to more sophisticated approaches that combine human intelligence with computer algorithms.

7.5 ACTIVITIES

1. Identify the distributed nature of services that you observe in a library that you think require moving from traditional text retrieval systems to WEB-based information retrieval.
2. Examine the information retrieval through a search engine. What criteria were used to produce and rank the results?

7.6 SELT ASSESSMENT QUESTIONS

1. How distributed users and platforms may benefit from XML for transferring data?
2. Why XML is important for information retrieval and sharing on the World Wide Web? What is its advantage on the HTML?

7.7 REFERENCES

Chowdhury, G. and Chowdhury, S., Information Sources and Searching on the World

Wide Web, London, Library Association Publishing, 2001.

Poulter, A., Hiom, D. and Tseng, G, The Library and Information Professionals' Guide to the Internet, 3rd edu, Library Association Publishing, 2000.

Bharat, K. and Henzinger, M. R., Improved Algorithms for Topic Distillation in a Hyperlinked Environment. In Croft, W. B., Moffat, A., Rijsbergen, C. J., Wilkinson, R. and Zobel, J. (eds), Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'98), ACM, New York, 1998, 10W111.

Unit–8

Intelligent Information Retrieval and Natural Language Processing and Applications in Information Retrieval

Complied by: **Dr. Munazza Jabeen**

Reviewed by: **1. Dr. Pervaiz Ahmad**
2. Dr. Muhammad Arif
3. Muhammad Jawwad

CONTENTS

Page #

INTRODUCTION

Where could intelligence be manifest in an information retrieval system? The inclusion of the user in the IR system, and the incorporation of interaction as a major process in IR, have some significant implications for how we might consider what would constitute intelligence in an IR system. For instance, under this view, the idea of the 'intelligent agent' seems untenable, at least in its most straightforward sense. That is, a program which takes a query as input, and returns documents as output, without affording the opportunity for judgment, modification and especially interaction with text, or with the program, is one which would not qualify as an IR system at all. Such a program would fail to know about the user's information problem (relying only upon the query, some poor representation of that problem), and would fail to incorporate that one process which is known to improve retrieval performance significantly, interaction (especially, but not exclusively, through relevance feedback). So, although we might say that the representation and comparison processes might be performed well, and even 'intelligently', the system would not perform intelligently (if by that, we mean well, or effectively).

Another point which this view of the IR system raises is that there are some processes in the IR system which cannot be performed by any other component than the user. Interaction is a joint process of user with the other components (also of the other components with one another), and judgment is a process that can only be performed by the user. Furthermore, although modification is something that can be done by the other components of the system with reference to modifying query or text representation, modification of understanding of the information problem is something that can realistically be done only by the user. Thus, the idea of the 'intelligent intermediary' as being the basis of intelligent IR, although perhaps necessary, is not sufficient to characterize the complete intelligent IR system. Similarly, the idea of good IR as being effective IR fails if all the intelligence is concentrated in only the built system since it thereby excludes the most significant aspect of effectiveness, the user's judgment of the comparison performance.

OBJECTIVES

After reading this unit you would be able to:

1. Understand cross language information retrieval
2. Learn machine translation through question-answering systems on web
3. Learn text mining for information extraction on Web
4. Identify information extraction methods

8.1 NATURAL LANGUAGE PROCESSING APPLICATIONS IN INFORMATION RETRIEVAL

The success of a retrieval system depends significantly on the way the documents in the database are represented. According to Blair, the process of representing documents for retrieval is fundamentally a linguistic process, and the most important problem relates to how language is used.' The linguistic problem of information retrieval can be viewed from two angles. From the point of view of input, the problem lies in the fact that the documents are represented (indexed) by people who have little or no idea of the terms that the users will select. From the user's point of view, the tasks for which the user wants information to determine what index terms they would like to use, and the terms actually used in the search will determine the nature of the final result. Traditional theories of document indexing, or representation consider two aspects: the context and the subject.

8.2 CROSS-LANGUAGE INFORMATION RETRIEVAL

The internet and the web have brought significant improvements in the way we create, look for and use information. However, information resources generated all over the world are not necessarily written in any one language, and users from the world over do not necessarily speak, and therefore seek information in, only one language. In other words, the language of an information resource and the language in which information is sought by a user may be different. So, how do such users get access to information that is in a different language? Cross-language information retrieval is a field of study that addresses this problem by using different techniques of natural language processing.

CLIR enables users to search and use information that is in a different language from their queries. Thus, a user may submit a query in a language that is different from the language of the target information resources; for example, the user may submit a query in Chinese and the CLIR system returns documents in English, or the user may submit a query in French and the system finds answers from

documents in English and then translates those answers into French. CLIR is often used interchangeably with terms such as ‘cross-lingual information retrieval’, ‘trans- lingual information retrieval’, ‘bilingual information retrieval’ and ‘multilingual information retrieval’.

8.2.1 EVERYDAY EXAMPLES OF CLIR

CLIR is a common feature of many search engines. For example, the Google Language Tools service allows users to enter a search in their own language; and Google will find other languages and translate the search results into the user’s language. Users can choose a language from a list of 42 different ones (Figure 20.1) and ask Google to search in automatically selected languages or in a preferred language.

8.3 MACHINE TRANSLATION

Although machine translation research began almost five decades ago, in the 1960s, the results were not very encouraging. However, as computers became more powerful, and new NLP techniques for machine translation were developed, better results began to emerge. With the proliferation of the web and digital libraries, multilingual information retrieval has become more common. There are two major challenges in this area: the recognition, manipulation, and display of multiple languages; and cross-language information search and retrieval. The former relates to the enabling technology that will allow users to access information in whatever language it is stored; while the latter implies permitting users to specify their information needs in their preferred language while retrieving information in whatever language it is stored. Text translation can take place at two levels: translation of the full text from one language to another for the purpose of search and retrieval.

8.4 QUESTION ANSWERING SYSTEMS

Conventional database management systems (DBMSs) are designed to find specific answers from structured databases, whereas conventional information retrieval systems are designed to find a set of documents where the user may find answers to their questions. In the case of DBMSs, structured data is retrieved in response to questions posed in highly structured query languages, such as SQL. As opposed to this, conventional information retrieval systems return unstructured texts that contain queries which are presented as a set of terms with optional structural connectives.

Question answering (QA) systems (or fact retrieval systems as they used to be called), although a type of information retrieval system, aim to provide answers to a question. According to Text Retrieval Conferences (TREC), ‘Question answering

systems return an actual answer, rather than a ranked list of documents, in response to a question' (http://trec.nist.gov/data/qa.html). However, although it is called a question answering system, a user does not necessarily have to put a query to the system exactly in the form of a question. In fact, QA systems aim to deal with a wide range of question types such as definition and meaning; fact-finding questions; what, how, and why types of questions; and so on.

8.5 QUESTION ANSWERING SYSTEMS ON THE WEB

START (http://start.csail.mit.edu) is the world's first QA system on the web. It was developed by Boris Katz and his associates of the Info Lab Group at the MIT Computer Science and Artificial Intelligence Laboratory and has been in existence since December 1993.

8.6 TEXT MINING

Text mining is a field of study that is designed to discover previously unknown information by automatically extracting information from a large collection of text documents. The extracted information may show relationships or patterns that are buried in the document collection. It can be used to analyze natural language documents about any subject, although there has been a lot of interest in biological sciences and security applications.

Text mining is like data mining, except it is designed to handle structured data from databases or XML files, working with unstructured or semi-structured data sets (such as e-mail, full-text documents, and HTML files). As a result, text mining is a much better solution for companies where large volumes of diverse information must be merged and managed. Text mining in bibliographic databases

8.7 INFORMATION EXTRACTION

Text mining and knowledge discovery have remained important areas of research over the past few years and a number of information science journals have published special issues reporting research on these topics (see for example, Benoit;Qin and Norton;Raghavan, Deogun and Server;Trybula; and Vicke@). Information extraction is a subset of knowledge discovery and text mining research that aims to extract useful bits of textual information from natural language texts. A variety of information extraction techniques are used, and the extracted information can be used for a number of purposes, for example to prepare a summary of texts, to populate databases to fill in slots in frames, and to identify keywords and phrase for information retrieval. Information extraction techniques are also used for classifying text items according to some predefined categories.

8.8 ACTIVITIES

1- Identify the processes of recognition, manipulation, and display of multiple languages in a machine translation.

2- Visit START (http://start.csail.mit.edu) the world's first QA system? Enlist the various features that make it an intelligent information retrieval system.

8.9 SELF ASSESSMENT QUESTIONS

1. How can text mining make an information retrieval system more intelligent for precise information extraction?
2. Explain why process of representing documents for retrieval is fundamentally a linguistic process?
3. How conventional information retrieval systems are different from the conventional database management systems?

8.10 REFERENCES

1 Blair, D. C., Language and Representation in Information Retrieval, New York, Elsevier, 1990.

Grosz, B. J., Weber, B. L. and Sparck Jones, K. (eds.), Readings in Natural Language Processing, New York, Morgan Kaufmann, 1986.

Obermeier, K. W, Natural Language Processing: an introductory look at some of the technology used in this area of artificial intelligence, Byte, 12, 1987, 225—32.

Jacobs, P. S. and Rau, L. F., Natural Language Techniques for Intelligent Information Retrieval. In Eleventh International Conference on Research and Development in Information Retrieval, New York, ACM, 1988, 85-99.

Chowdhury, G. G., Natural Language processing. In Cronin, B. (ed.), Annual Review of Information Science and Technology, 37, Medford, NJ, Information Today Inc., 2003, 51-89.

Haas, .S. W., Natural Language Processing: toward large-scale robust systems. In Williams, M. E. (ed.), Annual Review of Information Science and Technology, 31, Medford, NJ, Learned Information Inc. for the American Society for Information Science, 1996, 83-119.

Grishman, R., Natural Language Processing, Journal of the American Society for Information Science, 35, 1984, 291—6.

Warner, A. J., Natural Language Processing. In Williams, M. E. (ed.), Annual Review of Information Science and Technology, 22, Amsterdam, The Netherlands, Elsevier Science Publishers B. V. for the American Society for Information Science, 1987

UNIT–9

INFORMATION RETRIEVAL IN DIGITAL LIBRARIES AND TRENDS IN INFORMATION RETRIEVAL

Complied by: **Dr. Munazza Jabeen**

Reviewed by: **1. Dr. Pervaiz Ahmad**
2. Dr. Muhammad Arif
3. Dr. Amjid Khan

CONTENTS

Page #

INTRODUCTION

There are several definitions of digital libraries, many formulated during digital library research projects. Consequently, these definitions have been influenced by the people involved in the projects, by their understanding of the concept of libraries vis-a-vis electronic databases and by the nature of the research project. Borgman analyses several definitions of digital libraries and concludes that there are two major classes of definitions: those coming from digital library researchers — who in the US context are mostly computer scientists and engineers — and those coming from library and information professionals. The most comprehensive definition of a digital library, which emphasizes both the technical and the service aspects of digital libraries, was given during the March 1994 Workshop.

As we have already noted through the preceding units, information retrieval covers a vast area of study, and it is therefore difficult to keep track of the latest developments and consequently the trends in research in this field. Moreover, recent developments in web and digital libraries have brought a major revolution in information retrieval as many people encounter the web every day. On one hand information retrieval breaks down barriers of distance, users' characteristics, and the nature of digital content; on the other it is now a part of the everyday life of a much higher proportion of the population than in the past. The web has also raised people's expectations; many now expect that every bit of information can be obtained through the web easily and usually with no cost.

OBJECTIVES

After reading this unit you would be able to:

1. Identify information resources in a digital library through learning common features of digital libraries
2. Configure the basic design of a digital library
3. Learn the concepts of interoperable systems and its importance for digital libraries
4. Recognize new trends in information retrieval

9.1 INFORMATION RESOURCES IN DIGITAL LIBRARIES

Digital libraries provide access to different types of information sources in a variety of formats. For example, a digital library may contain simple metadata or catalogues of information resources, such as OPACs, or may contain the full text of documents, images, audio and video materials. The information resources may be available in different formats, and they may have been produced by using different types of hardware and software. For example, the text may be in MS Word, PDF or HTML format; images may be available in GIF or JPEG file formats; and so on. These information resources may reside on several different servers — local as well as remote — and they may have been indexed differently. All these issues make the information retrieval process very complex.

9.2 THE BASIC DESIGN OF A DIGITAL LIBRARY

Providing access to a variety of information resources residing on different computer systems in several parts of the world to several users of differing natures and needs is a major challenge for digital library designers. Digital libraries, especially hybrid libraries, aim to work as the 'one-stop shop' for all kinds of information resources. This involves a number of complex issues related to integration and seamlessness.

Figure 9.1 shows the basic design of a digital library. As this figure shows, users of a digital library may have access to a range of information resources and there are various modes of getting access to them. One possibility is for users to go through a custom-built interface that will allow them to select a particular type of resource, at which point the corresponding search interface opens for them to interact with it. Some digital libraries follow this model. A typical example is the Greenstone Digital Library, which allows users to select a particular type of resource or collection; then the interface of the corresponding resource is opened allowing the user to browse or search. The major problem of this model is that the user has to search or browse each collection separately.

Alternatively, users may choose one or more resources or collections and then formulate just one query, which is passed on to the various resources or collections by the digital library interface; results are brought back after the search is carried out. The user does not need to search the resources one by one, so this is a better approach from their perspective because they formulate only one search query and get results from all the different resources. A much better approach is now available in many digital libraries, which allows a cross-database search facility through one window without requiring the user to know about the specific collection or database to search. A typical example is the National Science Digital Library (NSDL; http://nsdl.org/g) in the USA, where the user can enter a search query to get results from a variety of information channels. However, technologically this approach is

more challenging, and several technical issues need to be considered in order to build this model.

e-journals		Online databases	Remote digital libraries		World wide web

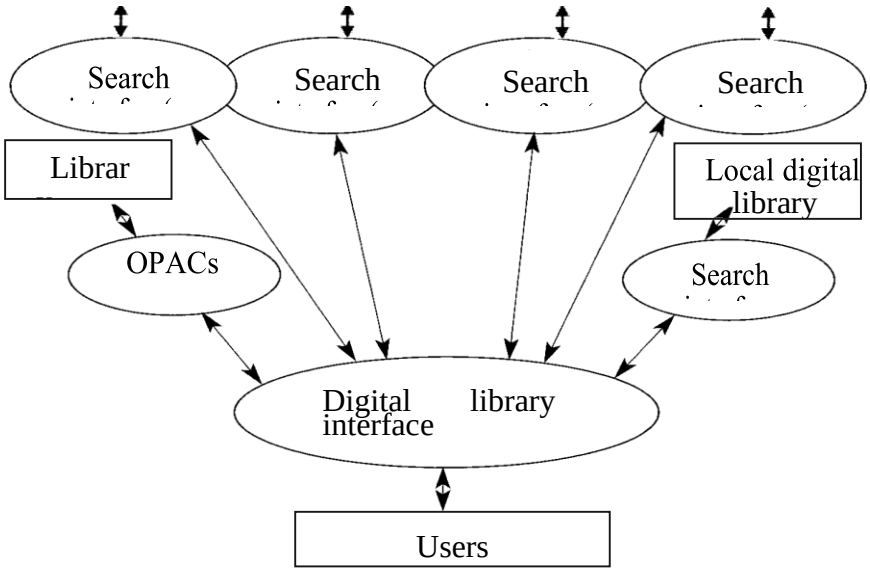

Figure 9.1 Conceptual designs of a digital library

9.3 INTEROPERABILITY

One of the major problems facing digital libraries is the issue of interoperability — how to get a wide variety of computing systems to work together and/or talk to one another for access to, and retrieval of, information. Interoperability and standardization are the most important considerations for digital library designers. There are different types of interoperability, such as systems interoperability, software interoperability or portability, semantic interoperability, linguistic interoperability, and so on. Interoperability among digital library systems can be achieved by several means, such as through adopting:

- common user interfaces
- uniform naming and identification systems
- standard formats for information resources
- standard metadata formats
- standard network protocols
- standard information retrieval protocols
- standard measures for authentication and security, and so on.

9.4 COMMON FEATURES OF DIGITAL LIBRARIES

- Meyyappan, Chowdhury and Foo reviewed the general features and Chowdhury, and Chowdhury reviewed the information retrieval features of some selected digital libraries. These are their main observations:
- Users can access the collections of a digital library by either browsing or searching.
- Although most digital libraries allow users to search the local digital library collections, some digital libraries provide facilities for federated search or search across several digital libraries.
- Boolean, proximity, and truncation search facilities are commonly available search options in digital libraries, although the operators vary. Some digital libraries provide options such as 'must also contain', 'or may contain', 'but not contain', 'should contain' and 'must contain' to activate a Boolean search.
- Keyword and phrase searches are common facilities of digital libraries, although the techniques for conducting a phrase search differ.
- Right truncation and wild card search facilities are common in many digital libraries, and a variety of operators, such as '%', '*', '@' and '?', are used for the purpose.
- Many digital libraries support proximity search differently. One of the options is to use proximity operators, but the operators vary, for instance 'Near', 'Nearby', 'Sentence', 'Paragraph' and so on.
- Most digital libraries allow users to conduct a search on specific fields. Although most digital libraries allow users to specify the maximum number of hits, the output is not always ranked, except in a few cases.
- In some cases, users can sort the results of a search using chosen keys. Usually, the system comes up with a brief output that can lead to the full records. However, in many cases an output format can be chosen by the user.

9.5 IFORMATION RETRIEVAL FEATURES OF SELECTED DIGITAL LIBRARIES

Information retrieval services are at the heart of digital libraries. Since a hybrid library can provide access to one or more of the information resources mentioned above, users may search each system separately using the search interface of each respective system. Alternatively, there may be a single search interface to allow users to conduct searches across all of the systems with just one query.

In the following sections we shall discuss the basic information retrieval facilities provided by some digital libraries. To facilitate our discussion, we have selected one digital library representing each one of the following loosely defined groups:

Type 1: fully fledged digital libraries that contain a variety of information distributed among a number of systems and platforms; however, users would like to use only one interface and get results from all the different systems by submitting only one query; although this poses a major challenge, some digital libraries provide such cross-database search facilities; NSDL in the USA and European digital library (www.europeana.eu/portal) are typical examples.

Type 2: digital libraries that provide access to some specific type of data, e.g., Music Australia, which provides access to music information, or PubMed, which provides access to health and related information.
Type 3: digital libraries that provide access to a variety of information resources, one at a time through a specific search interface, e.g., New Zealand Digital Library (NZDL).
Type 4: digital libraries that provide access to only one type of material, but allow a single or a multiple-site (federated) search, e.g. the Networked Digital Library of Theses and Dissertations
Type 5: digital libraries that provide access to all the different types of publications from a given publisher, e.g. ACM Portal.

9.6 NEED FOR DESIGN INTEGRATION

To design an effective and efficient information retrieval system for digital libraries, several layers of information system design need to be properly integrated. In short, the information retrieval system, in a digital library, aims to match the user requirements with the contents using the appropriate computer and networking technologies. However, different layers of work involving the organization and processing of information, user interfaces, networking, standards, and protocols, and so on, are involved in the process. All these different layers of work need to be properly integrated to develop a successful global digital library. Bates warns that: all layers of the system for accessing and displaying digital library information should be simultaneously designed with knowledge of what is going forward in the other layers. It takes only one wrongly placed layer to thwart all the clever work done at every other layer. For effective information retrieval to occur, all layers of a system must be designed to work together, and the people doing the designing must genuinely communicate.

9.7 NEED FOR INTERACTIVE QUESTION-ANSWERING SYSTEMS

Most information retrieval systems of today respond to user queries by retrieving documents or parts of one or more documents. However, ideally users would like to have specific answers to their questions. Building digital libraries that can provide answers in an interactive question-answering mode is a real challenge. It needs expertise from a number of fields including information retrieval, natural language processing, human—computer interactions, expert systems, and so on.
Several experimental question-answering systems are now being developed that aim to provide answers to natural language questions, as opposed to documents containing information related to the question. Such systems often use a variety of information extraction and retrieval operations using natural language processing tools and techniques to get the correct answer from the source texts. Breck et al. report on a question-answering system that uses techniques from knowledge representation, information retrieval and natural language processing. The authors claim that this combination enables domain independence and robustness in the face of text variability, both in the question and in the raw text documents used as knowledge sources. Research reported in the Question Answering track of TREC (Text Retrieval Conferences; http://trec.nist.gov) shows some interesting results. Now the experimental systems can provide answers to simple ‘who’ questions such

as ‘Who is the prime minister of Japan? and ‘when’ questions like ‘When did the Jurassic period end?’ The experimental systems work well as long as the query types recognized by the system have broad coverage, and the system can classify questions reasonably accurately. In TREC-the first QA track of TREC, the most accurate QA systems could answer more than two-thirds of the questions correctly. In the second QA track (TREC-9), the best performing QA system, the Falcon system from Southern Methodist University, was able to answer 65% of the questions. These results are quite impressive in a domain-independent question-answering environment. However, the questions were still simple in the first two QA tracks. In the future more complex questions requiring answers to be obtained from more than one documents will be handled by QA track researchers.

9.8 DISCUSSION

Information retrieval is one of the most fascinating, and yet challenging, areas in digital libraries. While years of research in text information retrieval are available to the researchers, the problems are multiplied by the volume, variety, format, and language of information resources coupled with the problems of the widely varying nature and requirements of users, and of information producers. Users of digital libraries should be familiar with the basics of information search techniques as well as with the information retrieval features of those systems that are accessible through the modern digital libraries. A number of working digital libraries provide reasonably good information retrieval features, especially for textual information retrieval. Results of experimental studies on multimedia and multilingual information retrieval are promising, and one can expect to see their applications in the future digital libraries.

9.9 TRENDS IN INFORMATION RETRIEVAL

One of the simplest applications of NLP in information retrieval has been in indexing based on some normalized or derived form of individual words occurring in input texts. Lexical level language analysis for the generation of automatic indexes was given much impetus with the arrival of machine-readable dictionaries. These dictionaries, such as the large ones available on CD-ROM (such as the Oxford English Dictionary), include a definition for each sense of a term, usually including syntactic category, morphology and perhaps semantic information such as restrictions on verb arguments and subject classifications. These dictionaries therefore help to produce more accurate descriptions of concepts in a text.

Humphrey and Miller produced a frame-based ‘index aid system’ as part of the Automated Classification and Retrieval Program (ACRP) undertaken at the Computer Science Branch of the US National Library of Medicine. Frames are used to represent indexable knowledge entities, in the domain of medical science, for processes, procedures, biological structures and chemical substances. The Topic Identification System (TIS) of Reuters, reported by Weinstein, uses artificial intelligence techniques to build rules that define the news topics; stories are categorized, indexed and entered into the database at the rate of three per second.

9.10 ACTIVITIES

1. Suppose your library needs to providing access to a variety of information resources residing on different computer systems in several parts of the world to a number of users differing natures and needs. What are the major considerations and design steps to you make as a digital library designer?
2. Visit the National Science Digital Library (NSDL; http://nsdl.org/g) in the USA. Describe its vital components in terms a digital library. How you think technologically this approach is more challenging and several technical issues need to be considered?

9.11 SELF ASSESSMENT QUESTIONS

1. Why common standard and metadata formats are vital in providing interoperable systems for digital libraries?
2. What is the role semantic interoperability in text mining and search engines?

9.12 REFERENCES

Borgman, C., What Are Digital Libraries? Competing visions, Information Processing and Management, 35 (3), 1999, 227W3.

Borgman, C., From Outenberg to the Global Information Ittfrastructure. access to information in the networked world, New York, ACM Press, 2000.

Gladney, H. H., Fox, E. A., Ahmed, Z., Asany, R., Belkin, N. J. and Zemankova, M., Digital Library: gross structure and requirements: report from a March 1994 Workshop, 1994, www.csdl.tamu.edu/DL94/paper/fox.html.

Oppenheim, C. and Smithson, D., What is the Hybrid Library?, Journal of Information Science, 25 (2), 1999, 97—112.

Pinfield, S., Eaton, J., Edwards, C., Russell, R., Wissenburg, A. and Wynne, P., Realizing the Hybrid Library, D-lib Magazine, October 1998, www.dlib.org/d1ib/october98/10pinfield.html.

Rusbridge, C., Towards the Hybrid Library, D-lib Magazine, 1998, www.dlib.org/d1ib/juIy98/rusbridge/07rusbridge.html.

Chowdhury, G. and Chowdhury, S., Introduction to Digital Libraries, London, Facet Publishing, 2003.

Arms, W., Digital libraries, Cambridge, MA, MIT Press, 2000.

HyLife, The Hybrid Library Toolkit: Interoperability, 2002, http://hy1ife.nun.ac.uk/toolkit/Interoperability.html.

Fox, E. A. and Sornil, O., Digital Libraries. In Baeza-Yates, R. and Ribeiro-Neto, B. (eds), Modern Information Retrieval, ACM Press, 1999, 415-32.